

UNIVERSIDAD DE COSTA RICA
FACULTAD DE DERECHO
ÁREA DE INVESTIGACIÓN



**TESIS PARA OPTAR POR EL GRADO ACADÉMICO DE
LICENCIATURA EN DERECHO**

*HACIA LA IMPLEMENTACIÓN DE SISTEMAS AUTOMATIZADOS DE DECISIÓN
BASADOS EN INTELIGENCIA ARTIFICIAL EN LA ADMINISTRACIÓN DE JUSTICIA
COSTARRICENSE: UN ANÁLISIS DESDE EL MARCO NORMATIVO DE LA UNIÓN
EUROPEA*

**KHEVIN ALBERTO SÁNCHEZ ZAMORA
B77143**

2025

San José, Costa Rica, 19 de febrero de 2024.

Universidad de Costa Rica.

Facultad de Derecho

Área de Investigación.

Dr. Tomás Federico Arias Castro.

Estimado Señor:

Reciba un cordial saludo. En mi calidad de director del Trabajo Final de Graduación titulado "*HACIA LA IMPLEMENTACIÓN DE SISTEMAS AUTOMATIZADOS DE DECISIÓN BASADOS EN INTELIGENCIA ARTIFICIAL EN LA ADMINISTRACIÓN DE JUSTICIA COSTARRICENSE: UN ANALISIS DESDE EL MARCO NORMATIVO DE LA UNIÓN EUROPEA*" elaborado por el estudiante Khevin Alberto Sánchez Zamora, carnet B77143; comunico que el anterior cuenta con mi visto bueno para ser defendido públicamente.

La presente investigación desarrolla un análisis exhaustivo sobre la implementación de sistemas de inteligencia artificial en la administración de justicia, partiendo de un examen pormenorizado de la evolución técnica de los modelos de lenguaje y su potencial aplicación en el ámbito judicial. En el Capítulo I, titulado "La Inteligencia Artificial y su Promesa para el Derecho", la investigación desarrolla una sólida exploración de la evolución histórica de la inteligencia artificial, sus fundamentos técnicos y sus perspectivas de aplicación en la administración de justicia. El análisis jurídico destaca los avances trascendentales suscitados por los grandes modelos de lenguaje, los cuales han superado hitos impensables como la aprobación de exámenes profesionales complejos. Empero, la tesis advierte, con rigor metodológico, que los innegables beneficios potenciales de la IA judicial en términos de eficiencia y coherencia deben ponderarse frente a los riesgos derivados de la opacidad algorítmica, los sesgos estadísticos y la potencial erosión de garantías fundamentales.

El Capítulo II, "La Regulación de la Inteligencia Artificial en la Administración de Justicia", se adentra en el examen exhaustivo del marco normativo pionero que la Unión Europea ha erigido para encauzar la irrupción de los sistemas de IA en los procesos judiciales. La investigación disecciona con pericia las principales iniciativas comunitarias, desde el Libro Blanco y la Carta Ética de la CEPEJ hasta el ambicioso Reglamento sobre IA (AI Act), el cual clasifica las aplicaciones judiciales como de "alto riesgo" y les impone obligaciones de transparencia, trazabilidad y control humano. Se evidencia la determinación europea de salvaguardar los derechos fundamentales mediante una regulación estricta pero no prohibitiva, en contraste con los enfoques más liberales de Estados Unidos o centralizados de China.

En el Capítulo III, "La Implementación de Sistemas de IA en la Administración de Justicia Costarricense", la tesis diagnostica con precisión los vacíos normativos e institucionales que obstaculizan una adopción jurídicamente garantista de la IA en los tribunales del país. Ante ello, se propone una hoja de ruta inspirada en la experiencia europea, que comprende reformas legales

para clasificar los sistemas por nivel de riesgo, la creación de una comisión especializada de supervisión, la formación obligatoria de jueces y funcionarios, y la implantación de auditorías algorítmicas permanentes. El corolario de la investigación apunta a que, dotándose de las herramientas jurídicas idóneas y desplegando la voluntad política necesaria, Costa Rica podría encauzar una transformación digital de su sistema judicial que combine los beneficios de la IA con la prudencia y humanidad características de una judicatura democrática, garante de los derechos fundamentales.

En consecuencia, considero que el trabajo cumple con todos los requisitos de fondo y forma exigidos por el Reglamento de Trabajos Finales de la Universidad de Costa Rica para optar por el grado de Licenciatura en Derecho.

OSCAR EDUARDO
GONZALEZ
CAMACHO (FIRMA)

Firmado digitalmente por
OSCAR EDUARDO
GONZALEZ CAMACHO
(FIRMA)
Fecha: 2025.02.19 16:27:59
-06'00'

Dr. Óscar Eduardo González Camacho.

5 de mayo de 2025

Magister
Tomás Federico Arias Castro
Director, Área de Investigación
Facultad de Derecho
Universidad de Costa Rica

Estimado profesor Arias:

Después de un cordial saludo, me sirvo indicar lo siguiente: como es de su conocimiento, formo parte del Comité Asesor del trabajo final de graduación (modalidad tesis) titulado "*Hacia la implementación de sistemas automatizados de decisión basados en inteligencia artificial en la administración de justicia costarricense: un análisis desde el marco normativo de la Unión Europea*", que sustenta el egresado Khevin Alberto Sánchez Zamora, carné B77143.

Por ello, en mi condición de Lector de la investigación, he revisado avances periódicos y dado una lectura integral final al documento aportado por la postulante. En ese sentido, le informo que este producto académico cumple con los requisitos de forma y fondo exigidos por la normativa universitaria; en consecuencia, le otorgo mi aprobación.

Con mis muestras de consideración y estima,

ANDREI
CAMBRONERO
TORRES (FIRMA)

Firmado digitalmente por
ANDREI CAMBRONERO
TORRES (FIRMA)
Fecha: 2025.05.05 12:36:43
-06'00'

Prof. Dr. Andrei Cambronero Torres
Lector del TFG

23 de abril 2025

Dr. Tomás Federico Arias Castro

Director

Área de Investigación

Facultad de Derecho

Universidad de Costa Rica

Estimado don Tomás:

Reciba un cordial saludo. Por la presente, hago de su conocimiento que, en mi condición de lector del trabajo final de graduación titulado "Hacia la implementación de sistemas automatizados de decisión basados en inteligencia artificial en la administración de justicia costarricense: un análisis desde el marco normativo de la Unión Europea", elaborado por el estudiante KHEVIN ALBERTO SÁNCHEZ ZAMORA, carné universitario número B77143, otorgo mi aval para que proceda con la defensa oral y pública del mismo, en virtud de que considero que cumple con los requisitos de forma y fondo exigidos por el Área de Investigación de la Facultad, conforme a las disposiciones del Reglamento General de los Trabajos Finales de Graduación vigente en la Universidad de Costa Rica.

Asimismo, estimo que se trata de un trabajo riguroso y bien desarrollado, que constituye un aporte pertinente al análisis jurídico de la inteligencia artificial aplicada al ámbito de la administración de justicia, con una lectura crítica del marco europeo como referencia normativa.

Quedo a su disposición para cualquier requerimiento adicional.

RONALD HIDALGO CUADRA (FIRMA)



Firmado digitalmente por RONALD HIDALGO CUADRA (FIRMA)
Fecha: 2025.04.23
14:01:04 -06'00'

Ronald Hidalgo Cuadra

Cédula 105950023

Heredia, 5 de mayo de 2025

A quien corresponda

Leí y corregí el Trabajo Final de Graduación denominado: *Hacia la implementación de sistemas automatizados de decisión basados en inteligencia artificial en la administración de justicia costarricense: un análisis desde el marco normativo de la Unión Europea*, elaborado por el estudiante Khevin Alberto Sánchez Zamora, para optar por el Grado de Licenciatura en Derecho.

Corregí el trabajo en aspectos tales como: construcción de párrafos, vicios del lenguaje que se trasladan a lo escrito, ortografía, puntuación y otros relacionados con el campo filológico y, desde ese punto de vista, considero que está listo para ser presentado como trabajo final de graduación, por cuanto cumple con los requisitos establecidos por la Universidad de Costa Rica.

Se suscribe cordialmente,



Carlos Díaz Chavarría
Carné colegiado: 409
Cédula: 4-0155-0936 Teléfono: 83 - 26 - 28 - 65
Escritor - Profesor universitario
Filólogo - Maestría en Literatura (UCR) – Maestría en Docencia Universitaria
Comentarista del programa PANORAMA (CANARA)
Presentador de la sección *Cuestiones del idioma* (Teletica – ~~Telered~~ – ~~Teleuno~~ TV)
Premio Internacional Pergamino de Honor al Mérito 2015
Premio Micrófono de Oro a la Excelencia Comunicativa 2015
Premio Mundial a la Excelencia Literaria 2019
Premio al Mejor Docente de Humanidades 2020
Premio Excelencia a la Trayectoria Profesional 2021
Premio Águila de Oro a la Excelencia Académica 2022
Premio al Mejor Docente de Humanidades 2024

Dedicatoria

A mi madre, Jenny Zamora, por la atención infinita dedicada a cada paso de este proceso, pero sobre todo, por haberme enseñado a leer, regalo que se convirtió en la herramienta que ha construido mi mundo y en el origen de todo lo aprendido, de este trabajo y de quien soy hoy

A mi abuela, Ana Torres, cuyo apoyo constante, tanto moral como práctico, fue mi sostén durante toda la universidad. Su decisión pionera de forjar un futuro profesional para nuestra familia no solo abrió puertas, sino que hizo posible que yo hoy recorra este camino.

A Fabiana Sánchez y Neythan Zamora, mis hermanos: por la complicidad de escucharme siempre, por los consejos directos y necesarios, y por el entusiasmo que me impulsaba. Este título es nuestro, un soporte más para el futuro que construimos juntos como familia.

A Tatiana Zamora, mi tía, quien sembró, quizás sin saberlo, la semilla de mi futuro profesional al ponerme, aun pequeño, frente a mi primera computadora. Su paciencia para enseñarme y su tolerancia con mis torpes inicios fueron el terreno fértil donde germinó mi curiosidad por la tecnología, llevándome directamente a la Inteligencia Artificial y a la realización de este proyecto.

A mi tío Yoxsy Chaves y mi primo Noah Sánchez, cuya compañía constante, tanto en las alegrías como en las adversidades, ha sido un ancla vital. Su presencia es parte integral del entorno familiar que nutrió este esfuerzo.

A mi padre, Alberto Sánchez, por el ejemplo de su propia formación que me alentó a seguir este camino, y por el apoyo constante que trascendió la distancia, asegurando las condiciones necesarias para llegar a esta meta.

Agradecimientos.

Mi más profundo reconocimiento es para el Dr. Óscar González Camacho, figura cardinal en mi desarrollo académico y profesional, quien ha conjugado con excepcional generosidad los roles de director, jefe, mentor y amigo. Le estoy inmensamente agradecido por la **atmósfera de absoluta confianza y libertad intelectual** que propició desde el inicio, permitiéndome trazar mi propio rumbo en esta investigación con autonomía. Su **fe inquebrantable en el valor de este proyecto** y su aliento constante fueron esenciales para navegar cada etapa. Admiro inmensamente su **agudeza intelectual y su sólida trayectoria como uno de los grandes juristas de este país**, siendo un privilegio difícil de expresar el haber contado con su guía experta y su perspectiva iluminadora. Por todo ello, que su nombre figure en esta tesis representa para mí un **honor insigne** y la validación más significativa de este esfuerzo culminado.

Extiendo mi agradecimiento a los lectores de esta tesis. A Don Ronald Cuadra, por su participación en el comité evaluador y el tiempo dedicado a la revisión de este trabajo. A Don Andrei Cambronero, le agradezco no solo sus comentarios precisos y valiosos sobre este texto, reflejo de una lectura atenta que impulsó mejoras significativas, sino también por una influencia decisiva que se remonta a mi primer año. Fue precisamente en la primera clase de Derecho que cursé en la Facultad, impartida por él, donde se despertó mi interés intelectual por esta disciplina, y su apoyo y afirmación tempranos fueron fundamentales para consolidar mi vocación en este campo. Su doble contribución, como formador inicial y lector crítico final, cierra un ciclo muy significativo en mi trayectoria académica.

Mi más profundo agradecimiento a la Universidad de Costa Rica, mi Alma Máter. En sus aulas recibí una formación de excelente calidad, teniendo el privilegio de aprender de juristas de gran renombre cuyo magisterio, en otras circunstancias, habría sido inalcanzable. La UCR fue el vehículo que me permitió trascender mis circunstancias de origen; para un joven nacido en una situación socioeconómica desafiante, la posibilidad de obtener un título profesional es un testimonio vivo de la poderosa labor de movilidad social que realiza esta benemérita institución. Este logro no sería ni remotamente imaginable sin su existencia y su compromiso.

A mis amigos de toda la vida: Christopher Bonilla (y a su novia Emilia Carvajal, amiga invaluable y apoyo constante en este proceso), José Pablo Alarcón, José Acuña, Pedro Acuña, Juan Pablo Barquero y Andy Gómez. Les agradezco el escucharme hablar de esto una y otra vez, pero sobre todo, por ser el escape perfecto con su amistad y buen humor, **elementos esenciales** para llegar hasta aquí.

Un agradecimiento de **especial relevancia** para Jessica Guillén. Como **interlocutora privilegiada y constante**, conoció en la mayor profundidad cada alegría, frustración, avance y retroceso inherentes a este largo proceso. Su **escucha atenta y reflexiva** no fue solo un alivio personal, sino un **pilar intelectual fundamental** que me permitió ordenar ideas y mantener el rumbo. Mi gratitud por ese acompañamiento es infinita.

Agradezco sinceramente a mis colegas Esteban Pérez y Mariana Montero. Estando un par de años adelante en la carrera, tuvieron la generosidad de atender cada una de mis **dudas** y **preocupaciones** sobre el complejo proceso de tesis. Su orientación fue invaluable, y es un orgullo profesional compartir hoy trinchera laboral con ellos.

Para Santiago Quirós, referente clave de mi paso por la Facultad y el amigo más cercano que conservo de esa etapa. Agradezco tu interés genuino en esta tesis y tu constante ánimo. Valoro enormemente tu amistad perdurable.

ÍNDICE DE LA INVESTIGACIÓN

Dedicatoria	vii
Agradecimientos.	viii
Resumen	xiii
Ficha Bibliográfica	xiv
Introducción	1
Justificación del Tema	1
Metodología	4
Contenido Capitular	8
Capítulo I. La Inteligencia Artificial y su Promesa para el Derecho: Un Análisis Minucioso de su Trayectoria, Técnicas Contemporáneas y Perspectivas de Aplicación en la Administración de Justicia	10
Preámbulo. La Interiorización de la Trayectoria Histórica de la IA como Mapa Esencial para Trazar su Incorporación en la Administración de Justicia.....	10
Sección I. En el Principio: los Primeros Pasos de la IA.....	23
1.1.1.- Conferencia de Dartmouth y el nacimiento del término “Inteligencia Artificial”	23
1.1.2- Pioneros y Visionarios: las Primeras Promesas de la IA.....	24
1.2.- El Recorrido Temporal de la IA: una Mirada Retrospectiva a su Evolución	31
1.2.1.- Años 50-60: la Época de la Lógica y las Reglas.....	31
1.2.2.- Años 70-80: el Reinado de los Sistemas Expertos	35
1.2.3.- Años 90-2000: el Resurgimiento de las Redes Neuronales	51
1.2.4.- 2010-Presente: la Era del Aprendizaje Profundo y los Large Language Models (LLMs)	62
2.- Los Large Language Models: Reescribiendo las Reglas del Juego	67
2.1.- GPT 3, GPT 3.5 y GPT-4: Buques Insignia de OpenAI.....	67
2.2.- Claude 2 de Anthropic y las Ventanas de Contexto	73
2.3.- Gemini 1.5 y Claude 3: Encontrando la Aguja en el Pajar	78
3.- Chain-of-Thought y Reasoning: O1 y Deepseek como Exponentes del Nuevo Paradigma en los Large Language Models	89
3.1.- Open AI O1	89
3.2- Deepseek R1	94
4.- Código Abierto vs Cerrado: La Transparencia en la Balanza	99
3.1.- Llama de Meta AI.....	100
3.2.- Mixtral de Mistral AI: Mixture of Experts	102

3.3.- El Dilema de la IA Legal: entre la Capacidad y la Explicabilidad.....	107
5.- La Vigencia Efímera del Estado del Arte Tecnológico: Una Guía Prospectiva.....	109
6.- El Ecosistema Legaltech: Contextualizando la IA como Vértice de la Transformación Tecnológica en el Derecho	112
7.- La IA en el Ámbito Judicial: Potencial y Desafíos.....	114
8. Epílogo del Capítulo I.....	140
Capítulo II. La Regulación de la Inteligencia Artificial en la Administración de Justicia: un Análisis Exhaustivo del Marco Normativo Europeo	146
2.1.- La Necesidad de un Marco Regulatorio para la IA en la Justicia.....	146
2.1.1.- Salvaguardia de los Principios Fundamentales del Estado de Derecho.....	146
2.1.2.- Transparencia y Rendición de Cuentas	148
2.1.3.- Consideraciones Éticas	149
2.2.- El Papel Pionero de la Unión Europea en la Gobernanza de la IA	150
2.3. The EU AI Act	150
2.3.1.- Naturaleza Jurídica y Ámbito de Aplicación	154
2.3.2.- Principios Fundamentales del Enfoque Regulatorio Europeo en Materia de Inteligencia Artificial	157
2.3.3.- Análisis del AI Act en relación con la Implementación de Sistemas de IA en la Administración de Justicia	189
2.4.- Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales y su Entorno	199
2.4.1.- Naturaleza y Alcance Jurídico	200
2.4.2.- Análisis de los Principios Rectores de la Carta	202
2.4.3.- Apéndices de la Carta Ética: Análisis y Contribuciones Esenciales	206
2.5.- Reglamento General de Protección de Datos (GDPR)	216
2.5.1.- Naturaleza Jurídica y Ámbito de Aplicación	216
2.5.2.- Principios de Protección de Datos y su Aplicación a la IA.....	218
2.5.3.- Bases Jurídicas para el Tratamiento y Decisiones Automatizadas	220
2.5.4.- Salvaguardias y Garantías del GDPR Aplicadas a la IA en el Ámbito Jurisdiccional...	223
2.6.- Resolución del Parlamento Europeo de 6 de octubre de 2021 sobre la Inteligencia Artificial en el Derecho Penal y su Utilización por las Autoridades Policiales y Judiciales en Asuntos Penales (2020/2016(INI))	229
2.7.- Libro Blanco sobre Inteligencia Artificial publicado por la Comisión Europea el 19 de febrero de 2020.....	242
2.8.- Análisis Comparativo con Marcos Regulatorios de otras Jurisdicciones Hegemónicas	248

2.8.1. Enfoque Regulatorio de Estados Unidos	250
2.8.2.- Modelo Regulatorio de la República Popular de China.....	257
2.8.3. Análisis Comparativo: Estados Unidos versus China en la Regulación de la IA	268
2.8.4. Conclusiones Finales: Convergencias y Divergencias respecto del Paradigma de la Unión Europea.....	272
2.9. Desafíos y Perspectivas Futuras de la Regulación de la IA en la Justicia Europea	279
2.9.1. Paradojas de la IA en la Justicia: Promesa y Cautela	280
2.9.2. La “Tercera Vía Garantista”: Fundamentos y Dilemas.....	280
2.9.3. El Temor a la “Justicia Codificada”: la Trampa de la Uniformización	284
2.9.4. El “Efecto Bruselas” y el Posible Coste Competitivo.....	284
2.9.5. Innovar sin Abdicar de la Prudencia: Cauces de Convergencia	285
2.9.6. La IA futura en la Justicia: Escenarios Prospectivos	286
2.9.7. El Rol de la Interpretabilidad Mecanicista y la Investigación de Vanguardia.....	287
2.9.8. Más Allá de Europa: la Justicia Algorítmica como Debate Global	287
2.9.9. Conclusiones: la Senda Europea hacia una Justicia Aumentada, No Reemplazada	288
Capítulo III: La Implementación de Sistemas de IA en la Administración de Justicia Costarricense: Diagnóstico, Desafíos y Propuestas desde la Experiencia Europea	289
3.1.- Estado Actual del Marco Normativo Costarricense.....	289
3.1.1.- Análisis del Marco Constitucional.....	289
3.1.2.- Normativa Vigente Aplicable.....	305
3.1.3.- Políticas y Directrices Institucionales	307
3.2.- Vacíos y Necesidades de Reforma	331
3.2.1. Brechas Normativas Identificadas.....	331
3.3.- Lecciones del Modelo Europeo.....	338
3.3.1. Principios y Garantías Adaptables	340
3.3.2. Mecanismos de Control y Supervisión	349
3.3.3. Estándares Técnicos y Operativos.....	353
3.4. Propuesta de Implementación.....	361
3.4.1. Reformas Legislativas Necesarias.....	361
3.4.2. Adecuación Institucional	369
3.4.3. Hoja de Ruta.....	384
3.4.4. Perspectivas del Análisis Económico del Derecho sobre la Implementación Propuesta	398
Conclusiones de la Investigación	402
Bibliografía	411

Resumen

Esta tesis aborda la integración de la inteligencia artificial (IA) en la administración de justicia, un campo de alta sensibilidad donde las promesas de eficiencia tecnológica confrontan la necesidad de salvaguardar derechos fundamentales. La **justificación** radica en la rápida expansión global de sistemas automatizados de decisión basados en IA (SADIA) y su incipiente exploración en Costa Rica, frente a la ausencia de un marco normativo específico que oriente su uso ético y seguro. La complejidad de la IA –su potencial para optimizar procesos, pero también para perpetuar sesgos o generar opacidad– exige una regulación que equilibre innovación y garantías constitucionales (independencia judicial, debido proceso, igualdad, privacidad), tomando como referente el pionero y garantista marco normativo de la Unión Europea (UE).

La **hipótesis** central sostiene que, si bien los SADIA ofrecen beneficios para la justicia costarricense (eficiencia, agilización), implican riesgos sustanciales (sesgos, erosión del rol judicial) que demandan una regulación específica, inspirada en principios constitucionales y en las mejores prácticas internacionales (particularmente las de la UE), como condición previa a una implementación a gran escala.

El **objetivo general** es analizar críticamente el marco normativo de la UE sobre IA judicial para extraer lineamientos aplicables a Costa Rica. La **metodología** empleada es cualitativa, basada en análisis documental y jurídico-comparado (centrado en la UE, contrastado con EE.UU. y China), culminando en una propuesta teórico-normativa para el contexto costarricense.

El **Capítulo I** contextualiza la evolución técnica de la IA (hasta LLMs como GPT-4, O1, R1 de Deepseek), reconociendo la rápida obsolescencia de los detalles específicos pero subrayando la **importancia estructural** de esta base para comprender las capacidades, limitaciones (opacidad, sesgos) y riesgos que la regulación debe abordar. Se remite a plataformas de seguimiento (Chatbot Arena, Scale LLM Leaderboard) para actualizaciones futuras.

El **Capítulo II** disecciona el marco normativo de la UE (AI Act, RGPD, Carta Ética CEPEJ), destacando su enfoque **basado en riesgo** (IA judicial como "alto riesgo"), sus principios rectores (antropocentrismo, transparencia, supervisión humana, no discriminación) y sus mecanismos de gobernanza multinivel.

El **Capítulo III** diagnostica la situación costarricense, identificando **vacíos normativos clave** (ausencia de ley específica sobre IA judicial, falta de clasificación de riesgo, lagunas procesales y de responsabilidad). Basado en las lecciones europeas, formula una **propuesta de implementación** que incluye: reformas legislativas (ley marco o sectorial), adecuación institucional (comisión especializada, unidad técnica, capacitación) y una hoja de ruta gradual (pilotos a corto plazo, escalamiento a mediano, consolidación a largo plazo).

Las **conclusiones principales** confirman la hipótesis: la IA judicial ofrece oportunidades y riesgos que exigen regulación. El modelo europeo, centrado en garantías, es un referente adaptable para Costa Rica. Se concluye que es **urgente y necesario** desarrollar un marco normativo costarricense que llene los vacíos actuales, asegurando la **supervisión humana efectiva, la transparencia algorítmica, la no discriminación y la rendición de cuentas**. El objetivo debe ser una "**justicia aumentada**" –donde la IA apoye y optimice la labor humana– y no una sustitución que deshumanice la función jurisdiccional, preservando así la legitimidad y los valores del Estado de Derecho costarricense en la era digital.

Ficha Bibliográfica

Sánchez Zamora, Khevin Alberto. *Hacia la implementación de sistemas automatizados de decisión basados en inteligencia artificial en la administración de justicia costarricense: un análisis desde el marco normativo de la Unión Europea*. Tesis de Licenciatura en Derecho, Facultad de Derecho, Universidad de Costa Rica, San José, Costa Rica, 2025.

Director: Óscar Eduardo González Camacho.

Palabras clave: Inteligencia Artificial, Administración de Justicia, Large Language Models, EU AI Act, Chain-of-Thought, Derecho Algorítmico, Sistemas Expertos, Redes Neuronales, Aprendizaje Profundo, Marco Regulatorio, Garantías Procesales, Justicia Digital, Derecho Tecnológico, Ética Judicial, Transparencia Algorítmica, Explicabilidad, GDPR, Protección de Datos, Decisiones Automatizadas, Debido Proceso Digital, Interpretabilidad Mecanicista, Derecho Comparado, Justicia Aumentada, Constitucionalidad, Innovación Judicial, Seguridad Jurídica, Automatización Judicial,

Democratización de la Justicia, Estado de Derecho Digital, Jurisdicción Electrónica, Reforma Judicial, Gobernanza Algorítmica, Derechos Fundamentales Digitales, Tutela Judicial Efectiva

Introducción

Justificación del Tema

La cuarta revolución industrial, impulsada por la inteligencia artificial (IA), llama con insistencia a las puertas de la administración de justicia, prometiendo una transformación radical en la forma en que se resuelven los conflictos y se garantizan los derechos. La implementación de sistemas automatizados de decisión basados en IA (SADIA) ya no es una mera hipótesis futurista, sino una realidad tecnológica en expansión que ofrece potenciales beneficios en términos de eficiencia, celeridad y acceso a la justicia, aspectos particularmente relevantes para sistemas judiciales, como el costarricense, que enfrentan desafíos persistentes de congestión y demanda ciudadana de respuestas oportunas. La capacidad de estos sistemas para procesar ingentes volúmenes de información, identificar patrones y agilizar tareas rutinarias podría, en principio, coadyuvar a la optimización de los recursos judiciales y a la reducción de la mora.

Sin embargo, esta incursión tecnológica no está exenta de sombras ominosas. La delegación, aun parcial, de funciones tradicionalmente reservadas al discernimiento humano en algoritmos cuya lógica interna puede ser opaca ("caja negra"), amenaza con erosionar pilares constitucionales irrenunciables en Costa Rica. La **independencia judicial** (Art. 154 Const.), concebida como la sujeción exclusiva del juez a la Constitución y la ley, podría verse comprometida si la "asistencia" algorítmica deriva en una influencia indebida o en una abdicación de la responsabilidad decisoria. Asimismo, la garantía inviolable del **debido proceso** (Arts. 39 y 41 Const.), que exige contradicción, defensa y motivación racional de las sentencias, se ve interpelada por sistemas cuyas inferencias no siempre son explicables o auditables por las partes.

Más aún, el riesgo de que los SADIA, entrenados con datos históricos, **perpetúen o incluso amplifiquen sesgos discriminatorios** preexistentes contra ciertos grupos sociales, representa una afrenta directa al **principio de igualdad ante la ley** (Art. 33 Const.) y a la prohibición de discriminación contraria a la dignidad humana. La posibilidad de una "justicia codificada", que aplique patrones estadísticos de manera inflexible sin atender a la singularidad del caso concreto,

choca frontalmente con la aspiración de una justicia material y equitativa. A ello se suman las preocupaciones sobre la **protección de datos personales** (Art. 24 Const. y Ley 8968), dada la ingente cantidad de información sensible que estos sistemas podrían procesar.

Ante este escenario de alto potencial y riesgo equivalente, Costa Rica se encuentra en una **encrucijada crítica**. Si bien existen iniciativas incipientes de digitalización y exploración de IA en el Poder Judicial, el país carece, hasta la fecha, de un **marco normativo específico y comprehensivo** que regule el desarrollo, la implementación y la supervisión de SADIA en el ámbito judicial. Este vacío normativo genera incertidumbre jurídica, dificulta la adopción responsable de tecnologías prometedoras y deja al sistema vulnerable a implementaciones apresuradas o carentes de las debidas garantías éticas y constitucionales.

Es precisamente en este contexto de necesidad regulatoria donde la mirada comparada adquiere una relevancia capital. La **Unión Europea**, consciente de la magnitud del desafío, ha emergido como un referente global al articular un **ecosistema regulatorio pionero y detallado** para la IA, con especial atención a sus aplicaciones de "alto riesgo", categoría en la que explícitamente se incluye la administración de justicia (EU AI Act). Este modelo europeo, fundamentado en principios como el antropocentrismo, la transparencia, la supervisión humana significativa, la gestión de riesgos y la no discriminación –y complementado por instrumentos como el RGPD y la Carta Ética de la CEPEJ–, ofrece no solo un catálogo de buenas practices, sino un **paradigma regulatorio maduro y profundamente alineado con la tradición garantista** que Costa Rica comparte. Su análisis crítico no busca una mera trasposición acrítica, sino extraer **lecciones valiosas y principios adaptables** que puedan informar el diseño de una regulación nacional robusta y contextualmente adecuada.

Por consiguiente, esta investigación se justifica por la **urgencia de dotar al ordenamiento jurídico costarricense de las herramientas conceptuales y normativas necesarias** para encauzar la inevitable integración de la IA en la administración de justicia. Analizar el enfoque europeo permite identificar soluciones probadas para mitigar los riesgos inherentes a los SADIA, asegurando que su implementación potencie la eficiencia y el acceso a la justicia **sin sacrificar los derechos fundamentales ni la integridad del Estado de Derecho**. Este estudio resulta, por tanto, indispensable no solo para la academia jurídica, sino también para los legisladores, los operadores

judiciales y la sociedad costarricense en su conjunto, al ofrecer una base informada para tomar decisiones trascendentales sobre el futuro tecnológico de la justicia en el país. Lejos de oponerse al progreso, se busca orientarlo por sendas que refuercen, y no debiliten, los pilares de una justicia democrática, transparente y centrada en la persona.

Objetivos

a.- General:

- Analizar críticamente el marco normativo de la Unión Europea relativo a la implementación de sistemas automatizados de decisión basados en Inteligencia Artificial, con el propósito de extraer lineamientos y buenas prácticas que sirvan de base para el desarrollo de una regulación en esta materia acorde con los valores y principios del sistema de administración de justicia costarricense.

b.- Específicos:

- I. Examinar los instrumentos jurídicos de la Unión Europea que establecen principios éticos y técnicos para el desarrollo, diseño y la aplicación de sistemas de inteligencia artificial en el ámbito de la administración de justicia.
- II. Identificar los principales beneficios y riesgos asociados a la implementación de sistemas automatizados de decisión en la resolución de conflictos judiciales, con base en las experiencias documentadas en los Estados Miembros de la Unión Europea.
- III. Formular una propuesta de lineamientos jurídicos y buenas prácticas dirigida a las autoridades costarricenses, que permita encauzar, responsablemente, un eventual proceso de incorporación de estas tecnologías en el sistema judicial nacional, en concordancia con los derechos fundamentales y el ordenamiento constitucional vigente.

Hipótesis

La implementación de sistemas automatizados de decisión basados en inteligencia artificial (SADIA) en la administración de justicia costarricense conlleva aspectos positivos y negativos, que deben sopesarse cuidadosamente.

Por un lado, estos sistemas pueden optimizar la gestión de casos sencillos y repetitivos, agilizando los procesos judiciales, reduciendo costos operativos y mejorando el acceso a la justicia. No obstante, también implican riesgos éticos y jurídicos que deben prevenirse, como la posible afectación del debido proceso, sesgos algorítmicos discriminatorios y la sustitución de la función esencial del juez.

Para un uso legítimo y adecuado de los SADIA en el ámbito judicial costarricense, es necesario establecer un marco normativo específico que regule sus aspectos técnicos, jurídicos y éticos. Dicha regulación debe basarse en los principios constitucionales nacionales y considerar las mejores prácticas desarrolladas en el ámbito internacional.

El marco regulatorio para la inteligencia artificial de la Unión Europea, contenido en instrumentos como la Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales o las directrices para la presentación electrónica de documentos judiciales, puede servir como modelo e inspiración valiosa para Costa Rica. Tomando en cuenta las particularidades del sistema jurídico nacional, se pueden adaptar los lineamientos europeos sobre derechos humanos, no discriminación, transparencia y control humano de los SADIA.

Así, una adecuada regulación de base ética y técnica puede potenciar los beneficios de la IA judicial, mitigando a la vez sus riesgos. Por ello, el desarrollo de un marco normativo propio conforme al Derecho costarricense resulta indispensable antes de implementar SADIA, para garantizar su uso legítimo en pro de la justicia.

Metodología

Esta investigación se articula metodológicamente con el propósito de abordar, con rigor sistemático y profundidad analítica, la compleja intersección entre la inteligencia artificial (IA), el marco normativo de la Unión Europea (UE) y su potencial implementación en la administración de justicia costarricense. La naturaleza del objeto de estudio —un fenómeno contemporáneo, tecnológicamente disruptivo y con profundas implicaciones ético-jurídicas— aconseja un abordaje que combine el análisis doctrinal y normativo con una reflexión propositiva informada.

A. Enfoque Metodológico

El **enfoque** adoptado para esta tesis es eminentemente **cualitativo**. Esta elección se fundamenta en la necesidad de explorar en profundidad las dimensiones conceptuales, normativas y axiológicas que subyacen a la regulación e implementación de la IA en el ámbito judicial. Un enfoque cuantitativo resultaría insuficiente para captar la complejidad de los debates sobre derechos fundamentales, ética algorítmica e independencia judicial que son centrales en este trabajo. Se busca comprender e interpretar los marcos regulatorios y las dinámicas institucionales, más que medir variables numéricas. Dentro de este marco cualitativo, la investigación se sustenta primordialmente en el **método documental**, complementado con un componente **teórico-propositivo**.

B. Materia y Delimitación del Objeto de Estudio

La **materia** central de la investigación es la viabilidad y las condiciones normativas para la implementación de sistemas automatizados de decisión basados en inteligencia artificial (SADIA) en la administración de justicia de Costa Rica. Para ello, se ha seleccionado, como principal referente comparativo, el **marco normativo desarrollado por la Unión Europea**.

La elección de la UE como paradigma de análisis no es fortuita, sino que responde a criterios estratégicos y epistemológicos precisos:

- **Vanguardia Regulatoria Global:** la UE se ha posicionado como líder global en la articulación de un marco jurídico comprehensivo y detallado para la IA (cristalizado en el Reglamento sobre IA o *AI Act*, junto con otros instrumentos como el RGPD y directrices éticas). Este corpus normativo representa, a la fecha, el intento más sistemático y ambicioso de regular la IA desde una perspectiva integral,
- **Enfoque Garantista ("Tercera Vía"):** frente a los modelos hegemónicos de Estados Unidos (predominantemente liberal, enfocado en la innovación y la autorregulación *ex post*) y China (centralista, orientado al control estatal y la estabilidad social), la UE ha optado por una "**tercera vía garantista**". Este enfoque prioriza la protección de los derechos fundamentales (dignidad humana, privacidad, no discriminación, debido proceso) y la preservación de los valores democráticos como eje rector de la regulación tecnológica. Esta primacía axiológica resuena profundamente con la tradición constitucional

costarricense, que también sitúa la dignidad y los derechos humanos en el centro del ordenamiento jurídico,

- **Pertinencia Comparativa frente a otras Regiones:** si bien existen otras jurisdicciones geográficamente más próximas a Costa Rica, como las latinoamericanas, que han explorado aplicaciones operativas interesantes de IA en la justicia (por ejemplo, el sistema Prometea en Argentina), se constata que, hasta la fecha, ninguna ha desarrollado un **marco normativo** con la robustez, sistematicidad y profundidad conceptual comparable al de la Unión Europea. Esta tesis busca extraer lecciones de un *modelo regulatorio* de vanguardia, más que replicar soluciones operativas aisladas cuya fundamentación normativa pueda ser menos elaborada o consolidada. El entramado jurídico de la UE ofrece una arquitectura más completa y madura para inspirar una regulación nacional.

C. Técnicas de Investigación

Para alcanzar los objetivos propuestos, se emplearán las siguientes **técnicas**:

- **Investigación Documental:** constituye la técnica principal. Se procederá a la recopilación, selección, el análisis crítico e interpretación sistemática de fuentes primarias y secundarias relevantes para el objeto de estudio,
- **Análisis Jurídico-Comparado:** se realizará una comparación sistemática entre el marco normativo europeo y el ordenamiento jurídico costarricense (constitucional, legal y reglamentario), identificando convergencias, divergencias, vacíos y posibles puntos de adaptación,
- **Estudio de Caso Disimilar (Contrastivo):** el análisis de los enfoques regulatorios de Estados Unidos y China se abordará como un **estudio de caso disimilar**. Su propósito no es la imitación directa, sino el contraste: al examinar modelos que difieren sustancialmente del europeo (uno más liberal, otro más estatalista), se busca resaltar por oposición las características distintivas, las fortalezas y los fundamentos del enfoque garantista de la UE. Este contraste abona a los objetivos de la tesis al permitir identificar los **parámetros virtuosos** (aquellos alineados con la protección de derechos y el Estado de Derecho) que deben inspirar la propuesta para Costa Rica, así como advertir sobre enfoques o prácticas

que resultarían incompatibles con nuestro marco constitucional. Ayuda a delimitar "lo deseable" frente a "lo evitable",

- **Elaboración Teórico-Propositiva:** sobre la base del análisis documental y comparativo, la tesis culminará con la formulación de una propuesta de lineamientos jurídicos y de adecuación institucional para Costa Rica. Esta propuesta no se deriva de investigación empírica de campo, sino de una construcción teórica razonada, fundamentada en las lecciones extraídas y en la aplicación de principios jurídicos al contexto nacional. Se trata de un **método propositivo de segundo nivel (teórico-propositivo)**, ya que no crea teoría jurídica desde cero, sino que aplica y adapta marcos conceptuales y normativos existentes (principalmente el europeo) para generar una propuesta concreta y contextualizada, orientada a la resolución de un problema jurídico-institucional identificado (la falta de regulación adecuada de la IA judicial en Costa Rica). Su fundamento radica en la capacidad de la investigación documental y comparada para informar una propuesta razonada y viable.

4. Fuentes de Investigación

Las **fuentes** empleadas se clasificarán en:

- **Fuentes Primarias:**
 - *Normativa Europea:* Reglamento sobre IA (AI Act), Reglamento General de Protección de Datos (RGPD), Carta Ética Europea sobre el Uso de la IA en la Justicia (CEPEJ), Resoluciones del Parlamento Europeo, Libros Blancos y Comunicaciones de la Comisión Europea. Se incluye explícitamente el análisis de instrumentos de *soft law*, dada su relevancia pre-legislativa e interpretativa en la configuración del marco europeo,
 - *Normativa Costarricense:* Constitución Política, Ley Orgánica del Poder Judicial, Códigos Procesales, Ley N.º 8968 de Protección de Datos, leyes sobre tecnología y firma digital, proyectos de ley en discusión sobre IA, políticas y directrices internas del Poder Judicial,

- *Jurisprudencia*: Decisiones relevantes de la Sala Constitucional de Costa Rica, del Tribunal de Justicia de la Unión Europea y, puntualmente, de tribunales de otras jurisdicciones (cuando ilustren principios o dilemas clave).
- **Fuentes Secundarias:**
 - *Doctrina Especializada*: libros, artículos en revistas científicas indexadas, videos especializados, capítulos de libros y tesis sobre IA, Derecho y Tecnología, Ética de la IA, Derecho Comparado, Derecho Constitucional y Procesal,
 - *Informes y Estudios*: publicaciones de organismos internacionales (ONU, UNESCO, OCDE, Consejo de Europa, BID), *think tanks*, centros de investigación académica y organizaciones de la sociedad civil que aborden la IA y la justicia.

El **criterio de selección** de la normativa y la literatura relevante se basará en: (i) **Pertinencia temática directa** con la IA en la administración de justicia; (ii) **Jerarquía normativa** (privilegiando Constitución, leyes y reglamentos vinculantes, pero sin descartar el *soft law* relevante); (iii) **Autoridad académica o institucional** de la fuente (priorizando publicaciones revisadas por pares, informes oficiales y doctrina reconocida) y (iv) **Actualidad** (dando preferencia a las fuentes más recientes, dada la rápida evolución tecnológica y regulatoria, sin perjuicio de recurrir a textos fundacionales cuando sea necesario para comprender la evolución conceptual).

Mediante la aplicación rigurosa de este enfoque y técnicas, se aspira a producir una investigación que no solo diagnostique la situación actual y analice el referente europeo, sino que, también, aporte lineamientos concretos y fundamentados para la implementación responsable y constitucionalmente adecuada de la inteligencia artificial en la administración de justicia costarricense.

Contenido Capitular

Para abordar de manera sistemática y coherente el complejo objeto de estudio, la presente investigación se articula en tres capítulos centrales, precedidos por esta introducción y seguidos por las conclusiones finales, configurando un recorrido lógico que avanza desde los fundamentos

técnicos y contextuales hasta el análisis normativo comparado y la formulación de propuestas concretas para el ordenamiento jurídico costarricense.

El **Capítulo I**, titulado “La Inteligencia Artificial y su Promesa para el Derecho: un Análisis Minucioso de su Trayectoria, Técnicas Contemporáneas y Perspectivas de Aplicación en la Administración de Justicia”, sienta las bases conceptuales indispensables para la comprensión del fenómeno bajo escrutinio. Se emprende un examen detallado de la evolución histórica de la inteligencia artificial, desde sus hitos fundacionales hasta el estado del arte actual, con especial énfasis en las capacidades y limitaciones de los modelos de lenguaje de gran escala (LLMs) y otras tecnologías relevantes. Se aborda, asimismo, la intrínseca volatilidad de este campo tecnológico y se justifica por qué, pese a ella, el entendimiento de estos fundamentos técnicos es crucial para dimensionar adecuadamente los desafíos regulatorios y éticos que la IA plantea al ámbito judicial. Este capítulo cumple, así, una función propedéutica esencial, proveyendo el sustrato técnico necesario para el análisis normativo subsiguiente.

Sobre esta base, el **Capítulo II**, denominado “La Regulación de la Inteligencia Artificial en la Administración de Justicia: un Análisis Exhaustivo del Marco Normativo Europeo”, se adentra en el núcleo del análisis comparado. Se examinan en profundidad los instrumentos jurídicos clave que conforman el ecosistema regulatorio de la Unión Europea en esta materia, destacando el Reglamento de IA (AI Act), el Reglamento General de Protección de Datos (RGPD) y la Carta Ética Europea sobre el Uso de la IA en los Sistemas Judiciales (CEPEJ). Se analizan los principios rectores de este modelo (enfoque basado en riesgo, antropocentrismo, transparencia, supervisión humana), los requisitos específicos impuestos a los sistemas de IA de "alto riesgo" – categoría en la que se incluye explícitamente la IA judicial– y los mecanismos de gobernanza y control establecidos. Este análisis permite identificar las estrategias normativas, las salvaguardas y las buenas prácticas desarrolladas por la UE para conciliar la innovación tecnológica con la protección de los derechos fundamentales, cumpliendo así con el objetivo central de extraer lecciones del referente europeo.

Finalmente, el **Capítulo III**, titulado “La Implementación de Sistemas de IA en la Administración de Justicia Costarricense: Diagnóstico, Desafíos y Propuestas desde la Experiencia Europea”, traslada el análisis al contexto nacional. Se realiza un diagnóstico del marco normativo

e institucional costarricense vigente (constitucional, legal, políticas internas del Poder Judicial, iniciativas legislativas en curso), identificando los vacíos, las brechas y las necesidades de reforma frente a los desafíos que plantea la IA. A partir de este diagnóstico y aplicando críticamente las lecciones extraídas del modelo europeo (Capítulo II), se formula una propuesta concreta y articulada. Esta incluye lineamientos para reformas legislativas necesarias, recomendaciones para la adecuación institucional (creación de órganos especializados, fortalecimiento de capacidades) y una hoja de ruta detallada con objetivos y acciones a corto, mediano y largo plazo. Este capítulo representa la culminación propositiva de la investigación, orientada a ofrecer soluciones viables y jurídicamente fundamentadas para Costa Rica.

El trabajo concluye con un apartado de **Conclusiones de la Investigación**, donde se recapitulan los principales hallazgos, se verifica la hipótesis planteada, se constata el cumplimiento de los objetivos y se ofrecen reflexiones finales sobre el futuro de la IA en la administración de justicia y la necesidad imperativa de un enfoque regulatorio garantista y centrado en la persona.

Capítulo I. La Inteligencia Artificial y su Promesa para el Derecho: Un Análisis Minucioso de su Trayectoria, Técnicas Contemporáneas y Perspectivas de Aplicación en la Administración de Justicia

Preámbulo. La Interiorización de la Trayectoria Histórica de la IA como Mapa Esencial para Trazar su Incorporación en la Administración de Justicia

En el alba de una nueva era, signada por el vertiginoso e inexorable avance de la inteligencia artificial, deviene imprescindible para todo jurista detenerse a reflexionar con ahínco y profundidad sobre las vastas y profundas implicaciones que esta tecnología revolucionaria ha de tener en el ámbito neurálgico de la administración de justicia.

Si bien *prima facie* podría parecer que una exposición circunstanciada sobre la evolución histórica y los intrincados fundamentos técnicos de la IA se aleja del núcleo esencial de esta tesis, sostengo que se trata de un paso ineludible e imperativo para comprender, en toda su magnitud y

complejidad, el fenómeno que nos ocupa y, además, para vislumbrar, con claridad meridiana, sus proyecciones futuras.

La inteligencia artificial, lejos de ser una moda efímera y pasajera, o una novedad surgida *ex nihilo* y sin antecedentes, es en realidad el fruto granado de décadas de investigación tenaz, de esfuerzos denodados y de avances exponenciales en múltiples campos del saber. Desde las primeras y seminales especulaciones teóricas de luminarias de la talla de Alan Turing y Claude Shannon¹ en los albores de la segunda mitad del siglo XX, pasando por los vaivenes entre el entusiasmo desbordante y la desilusión desencantada que signaron los años subsiguientes, hasta el renacimiento vigoroso experimentado en las últimas décadas gracias a la feliz confluencia de algoritmos más sofisticados, mayor poder computacional y vastos repositorios de datos, la IA ha recorrido un largo, fascinante y a menudo sinuoso camino.

Este recorrido no ha sido de ninguna manera lineal o previsible, sino que ha seguido una trayectoria de aceleración exponencial, un *crescendo* vertiginoso. En el núcleo mismo del vertiginoso progreso de la inteligencia artificial subyacen dos principios cardinales que han impulsado esta revolución tecnológica: la Ley de Moore y la Ley de Rendimientos Acelerados. Estos postulados, lejos de ser meras elucubraciones teóricas, han demostrado ser descripciones asombrosamente fidedignas de la realidad, corroboradas de manera reiterada por el curso del avance científico y tecnológico del último siglo.

La Ley de Moore, postulada por el cofundador de Intel, Gordon Moore, en su seminal artículo de 1965 *Cramming more components onto integrated circuits*², es uno de los principios rectores del progreso tecnológico en la era de la información. Esta ley enuncia que el número de transistores en un circuito integrado se duplica aproximadamente cada dos años, lo que conlleva un crecimiento exponencial de la potencia computacional.

En su artículo, Moore proporciona ejemplos concretos que ilustran la veracidad de su postulado y sus vastas implicaciones para el futuro de la electrónica. Señala que la complejidad de

¹Claude E. Shannon, "Programming a Computer for Playing Chess", The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 41, no. 314 (1950): 256-75, <https://doi.org/10.1080/14786445008521796>

²Gordon E. Moore, "Cramming More Components onto Integrated Circuits", Electronics 38, no. 8 (19 de abril de 1965), <https://www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf>

los circuitos integrados para un costo mínimo por componente había estado aumentando a un ritmo de aproximadamente un factor de dos por año, además predice que esta tendencia continuaría durante al menos la próxima década³.

Observa que, con las tolerancias existentes, ya era técnicamente factible construir transistores de alto rendimiento en centros separados por una distancia de apenas dos milésimas de pulgada. Para apreciar la magnitud de esta afirmación, es necesario tener presente que una milésima de pulgada equivale a 25.4 micrómetros, una dimensión unas 100 veces más delgada que un cabello humano. Que en un espacio tan minúsculo fuera posible construir no uno, sino varios transistores funcionales, es un testimonio del asombroso nivel de precisión y control que había alcanzado la industria de los semiconductores.

Pero Moore no se detiene ahí. Procede a ilustrar las implicaciones de esta capacidad de miniaturización. Señala que, en un cuadrado de dos milésimas de pulgada de lado, no solo podrían alojarse transistores, sino, también, otros componentes fundamentales de los circuitos electrónicos, como resistencias de varios kilos ohmios o incluso diodos. Esta densidad de componentes, argumenta Moore, permitiría albergar al menos 500 componentes en una pulgada lineal, o un asombroso cuarto de millón en una pulgada cuadrada.

La conclusión de este razonamiento es tan lógica como impactante: 65,000 componentes, una cantidad que en 1965 parecía astronómica, podrían para 1975 ocupar tan solo alrededor de un cuarto de pulgada cuadrada. Esta predicción resultó ser asombrosamente precisa.

³ Ibid, p. 83.

LA LEY DE MOORE EN FUNCIONAMIENTO

<i>Año</i>	<i>Transistores en el último chip de ordenador de Intel*</i>
1972	3500
1974	6000
1978	29000
1982	134000
1985	275000
1989	1200000
1993	3100000
1995	5500000
1997	7500000

Consumer Electronics Manufacturers Association

FIGURA 1: Manifestación de la Ley de Moore de los años 1972-1997⁴.

En su libro *The Age of Spiritual Machines* (La era de las máquinas espirituales), el futurista y tecnólogo Ray Kurzweil ofrece una explicación lúcida de las implicaciones de esta proyección. Según Kurzweil, este aumento en la densidad de transistores tiene dos efectos significativos: duplica la cantidad de elementos en un chip y duplica la velocidad de procesamiento. Lo más destacable es que este avance se logra manteniendo constante el costo de producción de los circuitos integrados.

⁴ Ray Kurzweil, *The Age of Spiritual Machines* (Barcelona: Editorial Planeta, S.A., 1999; reimpresión exclusiva para México, México, D.F.: Editorial Planeta Mexicana, S.A. de C.V., 2000), 37.

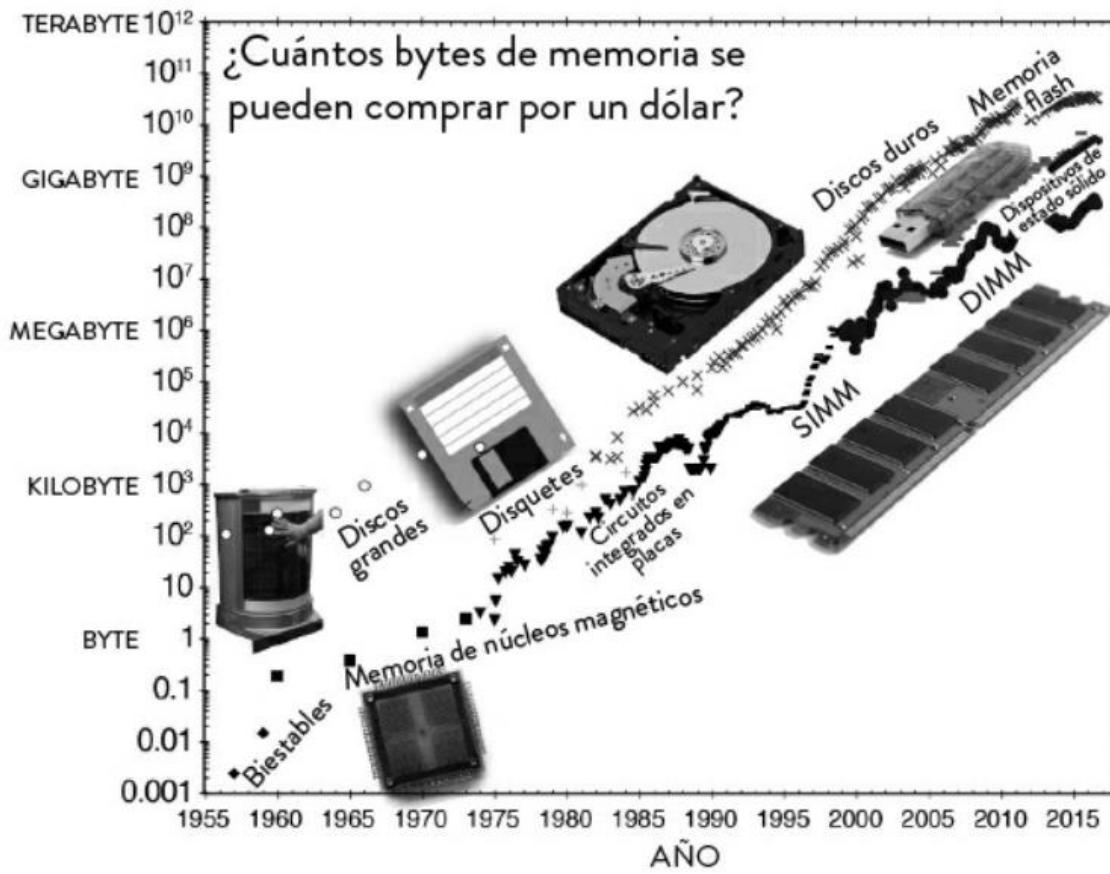


FIGURA 2: A lo largo de las seis décadas pasadas, el precio de la memoria de ordenador se ha reducido a la mitad aproximadamente cada dos años, lo que equivale a que sea unas mil veces más barata cada veinte años.⁵

La consecuencia de esta tendencia es que **cada dos años es posible obtener el doble de circuitos funcionando al doble de velocidad por el mismo precio**. Para muchas aplicaciones, esto representa una cuadruplicación del valor, ya que se obtiene un rendimiento exponencialmente mayor sin un aumento proporcional en el costo⁶. Esta observación se aplica a todo tipo de circuitos, desde chips de memoria hasta procesadores informáticos, lo cual ha impulsado un crecimiento acelerado en la capacidad de cómputo durante las últimas décadas.

Un ejemplo paradigmático de la Ley de Moore en acción es la evolución de las computadoras personales. En los albores de la era informática, las computadoras eran dispositivos

⁵ Max Tegmark, *Vida 3.0: Ser humano en la era de la Inteligencia Artificial*. (España: Editorial Taurus, 2018), p. 69. Recuperado de: ([PDF](#)) Vida 3. | Doroteo Arango - Academia.edu

⁶ Ibid., p. 40.

enormes, costosos y de acceso restringido. La IBM 650, lanzada en 1953, ocupaba una habitación entera⁷ y costaba 500.000 dólares de la época, pudiendo alquilarse por 3.500 dólares mensuales⁸. Sin embargo, gracias a los avances en la miniaturización de los componentes electrónicos impulsados por la Ley de Moore, las computadoras se han vuelto cada vez más pequeñas, potentes y asequibles.

La introducción de la microcomputadora Altair 8800 en 1975, considerada la chispa que encendió la revolución de las computadoras personales, representó un hito en este proceso de democratización tecnológica⁹. Con un precio de unos 395 dólares, la Altair 8800 puso la informática al alcance de un público más amplio. En las décadas siguientes, la Ley de Moore impulsó una carrera frenética de innovación que llevó al desarrollo de computadoras personales cada vez más potentes y asequibles, como la Apple II, la IBM PC y la Macintosh.

Hoy, un smartphone de gama media tiene una capacidad de procesamiento varios órdenes de magnitud superior a la de las supercomputadoras de hace apenas unas décadas, y a una fracción del costo. El iPhone 12, por ejemplo, cuenta con un procesador A14 Bionic con 11,8 billones de transistores, una cifra que habría sido inconcebible en los inicios de la era informática¹⁰. Esta miniaturización y aumento de la potencia de cómputo han permitido que la informática se integre en casi todos los aspectos de nuestra vida cotidiana, desde la comunicación y el entretenimiento hasta la educación y el trabajo.

⁷ "Overview", IBM Heritage: The IBM 650, acceso el 29 de marzo de 2024, <https://www.ibm.com/heritage/ibm650/>.

⁸ JJ Velasco, "Historia de la tecnología: IBM 650, el primer modelo fabricado en serie por el gigante azul", Hipertextual, 29 de agosto de 2011, <https://hipertextual.com/2011/08/ibm-650-primera-produccion-serie>

⁹ National Museum of American History, "Altair 8800 Microcomputer." Smithsonian, https://americanhistory.si.edu/collections/nmah_334396

¹⁰ Apple, "Apple Unveils All-New iPad Air with A14 Bionic, Apple's Most Advanced Chip," Apple Newsroom (15 de setiembre de 2020). Recuperado de: <https://www.apple.com/newsroom/2020/09/apple-unveils-all-new-ipad-air-with-a14-bionic-apples-most-advanced-chip/>



FIGURA 3: Desde 1900, el coste de la computación se ha reducido a la mitad aproximadamente cada dos años. La gráfica muestra la potencia de computación medida en operaciones de punto flotante por segundo (FLOPS) que pueden comprarse por mil dólares.¹¹

Estos ejemplos ilustran cómo la Ley de Moore ha sido un motor fundamental de la revolución digital que ha transformado nuestra sociedad en las últimas décadas. Desde las computadoras personales y los smartphones hasta la inteligencia artificial y la robótica, los avances en la miniaturización y el aumento de la potencia de cómputo han abierto nuevos horizontes y han planteado desafíos y oportunidades sin precedentes a una velocidad abismal.

No obstante, Kurzweil advierte que este enfoque convencional de reducción de tamaño eventualmente llegará a su límite. Más pronto que tarde, los transistores alcanzarán un tamaño de solo unos pocos átomos de espesor y la estrategia de miniaturización ya no será viable. En ese

¹¹ Ibid., p. 80.

punto, será necesario explorar nuevos paradigmas y nuevas tecnologías para continuar el avance de la capacidad computacional.

Cabría indagar sobre los motivos que subyacen a esta tendencia de progresión geométrica, no circunscrita únicamente al fenómeno de la miniaturización de los transistores (Figura 1), sino que se manifiesta de manera omnipresente en el vasto dominio de la computación en su conjunto (Figura 3), en la esfera de la memoria (Figura 2), así como en un sinnúmero de otras tecnologías que abarcan desde la secuenciación del genoma, hasta la predicción de la estructura tridimensional de las proteínas a partir de su secuencia de aminoácidos¹².

Esta duplicación incesante, en la que la tecnología exhibe lo que en términos matemáticos se denomina un crecimiento exponencial al multiplicar su potencia a intervalos periódicos, ha sido bautizada por Ray Kurzweil con el título de *Ley de Rendimientos Acelerados*:

“Ley de la Aceleración de Resultados: A medida que el orden crece en forma exponencial, el tiempo se acelera exponencialmente (es decir que, con el paso del tiempo, el intervalo entre acontecimientos destacados se acorta)”.¹³

La Ley de la Aceleración de Resultados o de Rendimientos Acelerados, postula que **el ritmo del avance e innovación tecnológica se incrementa de forma exponencial**, conduciendo a avances cada vez más rápidos y profundos en la capacidad de procesamiento de la información.

En el núcleo mismo de este postulado, palpita la noción de que el crecimiento exponencial no es un fenómeno aislado, una mera curiosidad matemática confinada a los estrechos límites de un campo específico, sino que forma parte de un continuo evolutivo que ha moldeado el cosmos desde sus albores, tejiendo un intrincado tapiz de cambio y transformación. Kurzweil argumenta que este patrón se manifiesta en una miríada de procesos, desde la expansión del universo hasta el delicado despliegue de la vida en nuestro planeta. Sin embargo, es en el reino de la tecnología donde esta aceleración alcanza su máxima expresión, impulsada por la sutil y poderosa

¹² John Jumper et al., "Predicción de estructuras de proteínas altamente precisa con AlphaFold", Nature 596, n.º 7873, 26 de agosto de 2021: 583-589. Recuperado de: <https://www.nature.com/articles/s41586-021-03819-2>

¹³ Kurzweil, p. 50.

retroalimentación entre la innovación y su aplicación, en un baile eterno de causa y efecto, de creación y consecuencia.

Así, aunque la Ley de Moore, que predice la duplicación de la densidad de transistores en los circuitos integrados cada dos años, ha sido durante décadas el paradigma dominante en la industria de la computación; esta ley no es más que un eslabón en una cadena mucho más amplia y exponencial. Aun cuando lleguemos al límite atómico de aquella, la innovación exponencial que estamos viviendo en el acontecer tecnológico promete expandir los horizontes de la capacidad computacional.

Sin embargo, es importante destacar que este crecimiento exponencial no se produce de forma lineal o predecible. Al igual que en la parábola del inventor del ajedrez y el emperador chino¹⁴, los avances más significativos suelen concentrarse en la segunda mitad del tablero, cuando la duplicación de la capacidad computacional comienza a manifestarse de manera más evidente. Estamos, por tanto, en un momento crítico de la historia, un punto de inflexión en el devenir de nuestra especie, en el que la aceleración de la innovación nos sitúa en el umbral de cambios sin precedentes, de transformaciones que desafían nuestra comprensión y ponen a prueba nuestra capacidad de adaptación.

Ante este horizonte que se despliega ante nosotros, la pregunta que debemos plantearnos los juristas no es si deberíamos o no implementar la IA en la administración de justicia, sino cuándo y cómo hacerlo de manera responsable y fructífera. Tal como sucedió con la adopción de Internet, las herramientas telemáticas y otras tecnologías disruptivas que han transformado radicalmente nuestra sociedad, la integración de la IA en los sistemas judiciales es una cuestión de tiempo. Y si

¹⁴ Existe un cuento popular que dice que hace mucho tiempo, un inventor presentó el juego de ajedrez a un emperador chino. El emperador quedó tan impresionado por la belleza y complejidad del juego que le ofreció al inventor cualquier recompensa que pidiera. El inventor, de manera inteligente, pidió algo que parecía simple pero que escondía una complejidad matemática enorme: solicitó que se le diera un grano de arroz por la primera casilla del tablero de ajedrez, dos por la segunda, cuatro por la tercera, y así sucesivamente, doblando la cantidad de granos de arroz en cada una de las 64 casillas del tablero. Al principio, el emperador se río de la humildad de la petición, pero pronto se dio cuenta de que el total de granos de arroz que el inventor había pedido era astronómicamente alto. La cantidad de arroz requerida se duplicaba con cada casilla, resultando en un número final tan grande que superaba con creces la producción de arroz de todo el reino, e incluso, según algunas versiones de la historia, más de lo que se podía encontrar en todo el mundo en ese momento. El número total de granos de arroz que se habrían necesitado para llenar las 64 casillas del tablero de ajedrez es $2^{64} - 1$, que es igual a 18,446,744,073,709,551,615 granos de arroz. Esto sirve como una poderosa ilustración del concepto matemático del crecimiento exponencial, mostrando cómo algo que comienza siendo pequeño puede llegar a ser enormemente grande con el tiempo si sigue duplicándose.

extrapolamos con rigor las tendencias exponenciales que hemos atestiguado en las últimas décadas, ese momento, ese punto de inflexión, puede estar más cerca de lo que muchos creen o quisieran creer.

En este contexto, las recientes declaraciones de Darío Amodei, una figura líder en este campo, merecen una consideración detenida y profunda por parte del lector. Amodei, fundador y CEO de Anthropic, una empresa en la vanguardia del desarrollo de sistemas de IA seguros e interpretables, atesora una trayectoria impresionante que lo posiciona como una voz autorizada para dilucidar los desafíos y oportunidades que esta tecnología transformadora plantea.

Antes de fundar Anthropic, Amodei se desempeñó como vicepresidente de Investigación en OpenAI, donde lideró el desarrollo de modelos de lenguaje de gran escala como GPT-2 y GPT-3, avances que han redefinido las fronteras de la IA en el procesamiento del lenguaje natural. Asimismo, es coinventor del paradigma de aprendizaje por refuerzo a partir de retroalimentación humana, una técnica que busca alinear los sistemas de IA con los valores y preferencias humanas. Su experiencia se nutre de su paso por Google Brain como investigador científico senior y de una sólida formación académica, con un doctorado en biofísica de la Universidad de Princeton como becario Hertz y estudios postdoctorales en la Escuela de Medicina de la Universidad de Stanford. Es desde este lugar de experticia que Amodei aborda la naturaleza de las denominadas "leyes de escalabilidad" – referidas líneas arriba - que parecen gobernar el ritmo de progreso de la IA:

"ENTREVISTADOR: Entonces, ¿cuál es tu perspectiva sobre la línea de tiempo en la que lograremos AGI, también conocida como inteligencia artificial poderosa o extremadamente útil? (...) en términos de pura inteligencia, aquella podría ser más inteligente que un ganador del Premio Nobel en todas las disciplinas relevantes (...).

DARIO: Si extrapolamos las curvas que hemos tenido hasta ahora, si dices, bueno, no lo sé, estamos empezando a llegar al nivel de doctorado y el año pasado estábamos en el nivel de pregrado y el año anterior estábamos en el nivel de un estudiante de secundaria. De nuevo, se puede discutir en qué tareas y para qué, todavía nos faltan modalidades, pero esas se están agregando, como se agregó el uso de la computadora, como se agregó la imagen, como se agregó la generación de imágenes. Si simplemente miras la velocidad a la que estas capacidades están aumentando, te hace pensar que llegaremos allí en 2026 o

2027. De nuevo, muchas cosas podrían descarrilarlo: podríamos quedarnos sin datos, podría no ser posible escalar los clústeres tanto como queremos, tal vez Taiwán sea destruido o algo así, y luego no podríamos producir tantas GPUs como queremos. Hay todo tipo de cosas que podrían descarrilar todo el proceso, así que no creo totalmente en la extrapolación en línea recta, pero si crees en la extrapolación en línea recta, llegaremos ahí en 2026 o 2027.

Creo que lo más probable es que haya algún retraso leve en relación con eso. No sé cuál sería ese retraso, pero creo que podría suceder a tiempo, creo que podría haber un retraso leve. Todavía existen escenarios en los que no ocurre en cien años, pero el número de esos escenarios está disminuyendo rápidamente. Nos estamos quedando sin razones verdaderamente convincentes o barreras realmente sólidas que expliquen por qué esto no sucederá en los próximos años. Había muchas más en 2020, aunque mi suposición, mi coronada en ese momento, era que superaríamos todas esas barreras. Así que, como alguien que ha visto la mayoría de las barreras despejadas, sospecho, mi coronada es que el resto no nos detendrá. Pero bueno, al final del día, no quiero representar esto como una predicción científica. A la gente le gusta llamarlas "leyes de escalabilidad", pero es un término incorrecto. La ley de Moore y las leyes de escalabilidad no son leyes del universo, son regularidades empíricas".¹⁵

Hay múltiples aristas relevantes de lo allí transcrito, que abonan a la impresión que este prologo busca en el lector. En primer lugar, Amodei advierte con lucidez que las “leyes” aquí expuestas son, en realidad, regularidades empíricas. Esta distinción, lejos de ser meramente semántica, tiene profundas implicaciones para proyectar el futuro de la IA. Mientras que las leyes de la naturaleza, como las de la física, son prescriptivas y deterministas, las regularidades empíricas son patrones observados, sujetos a variación y disruptión. Que hasta ahora el progreso de la IA haya seguido estas “leyes” no garantiza que así será indefinidamente. No obstante, Amodei se inclina a apostar por su continuidad, al menos en el corto y mediano plazo, dado su asombroso poder predictivo hasta la fecha.

¹⁵ Lex Fridman, “Dario Amodei: Anthropic CEO on Claude, AGI & the Future of AI & Humanity”, episodio 452 de Lex Fridman Podcast, 54:49, video de YouTube, publicado el 11 de noviembre de 2024, <https://www.youtube.com/watch?v=ugvHCXCOmm4&t=3581s>

Otro punto que no podemos perder de vista es el horizonte temporal que propone (2026-2017) para el horizonte temporal para el advenimiento de sistemas de inteligencia artificial de nivel humano, comúnmente denominados “AGI”. Esta proyección, aunque admitidamente especulativa, plantea un escenario que el derecho no puede permitirse ignorar. La posibilidad de que en un horizonte de apenas tres a cuatro años dispongamos de sistemas de IA capaces de igualar o superar el intelecto humano en prácticamente todos los dominios cognitivos, representa un desafío sin precedentes para nuestros marcos normativos y éticos. Si bien Amodei reconoce la posibilidad de retrasos y obstáculos, su convicción es clara: estamos agotando rápidamente las razones verdaderamente convincentes por las que esto no sucederá en los próximos años.

Esta apuesta tiene consecuencias trascendentales, particularmente en lo que ataña a los riesgos inminentes de la IA. Amodei advierte que, si bien los sistemas actuales aún no son lo suficientemente poderosos para presentar catástrofes serias, el caso de preocupación, el caso de riesgo, es lo suficientemente fuerte como para que debamos actuar ahora:

“DARIO: Creo que tenemos un dilema interesante con los sistemas de inteligencia artificial: aún no son lo suficientemente poderosos como para representar catástrofes serias. No sé si alguna vez llegarán a causar esas catástrofes; es posible que no lo hagan. Sin embargo, el argumento para preocuparse y el caso para considerar los riesgos son lo suficientemente sólidos como para que actuemos ahora. Además, están mejorando muy, muy rápido. Como mencioné en mi testimonio ante el Senado, podríamos enfrentarnos a riesgos biológicos graves en un plazo de dos a tres años, y eso fue hace aproximadamente un año. Las cosas han avanzado rápidamente desde entonces. Así que nos encontramos en esta situación en la que es sorprendentemente difícil abordar estos riesgos porque no están presentes hoy; no existen todavía, son como fantasmas, pero se acercan a nosotros a una velocidad impresionante debido a las rápidas mejoras de los modelos. Entonces, ¿cómo lidiamos con algo que no está aquí hoy, que no existe aún, pero que se aproxima tan rápidamente?”¹⁶

Aquí radica el dilema central: ¿cómo abordar riesgos que, aunque aún no se han materializado, se acercan a una velocidad asombrosa? Amodei compara estos riesgos con

¹⁶ Ibid.

fantasmas: intangibles pero ineludibles, especulativos, pero potencialmente catastróficos. La dificultad de calibrarlos y anticiparlos no los hace menos apremiantes; por el contrario, exige una mayor proactividad y diligencia por parte de quienes tenemos la responsabilidad de velar por el bien común.

Ante este panorama, la responsabilidad de la comunidad jurídica es ineludible. Recae sobre nosotros el deber de adelantarnos a estos desafíos, de desarrollar marcos normativos y herramientas conceptuales para encauzar el desarrollo de la IA de manera segura y beneficiosa para la sociedad. No podemos darnos el lujo de ser reactivos, de esperar pasivamente a que estos riesgos se materialicen para recién entonces intentar regularlos de manera improvisada y apresurada.

El imperativo ético es claro: actuar, y actuar ahora, por más especulativos que puedan parecer los riesgos desde la comodidad del presente. La incertidumbre no nos exime de esta responsabilidad; por el contrario, exige de nosotros una mayor diligencia y proactividad.

No podemos darnos el lujo de ser reactivos, de esperar pasivamente a que la IA irrumpa de manera descontrolada en nuestros tribunales para recién entonces, a *posteriori*, intentar regularla de manera improvisada y apresurada. Debemos ser proactivos, debemos forjar desde ahora, con visión de largo plazo, los marcos éticos y legales que aseguren que esta poderosa herramienta se utilice de manera justa, transparente y beneficiosa para el conjunto de la sociedad.

Pero, para hacerlo, para estar a la altura de este desafío mayúsculo, necesitamos primero comprender en profundidad qué es la IA, de dónde viene, cuáles son sus principios rectores y hacia dónde puede llevarnos. Necesitamos sumergirnos en su rica y compleja evolución histórica, en las ideas visionarias y los hitos técnicos que han marcado su desarrollo, en las mentes preclaras que han contribuido a forjarla. Solo entonces, con ese bagaje intelectual, con esa comprensión matizada y profunda, podremos entablar un diálogo fructífero y responsable sobre su aplicación en un ámbito tan neurálgico y sensible como la administración de justicia.

No se trata, pues, de adquirir conocimientos técnicos por mero diletantismo intelectual, por una vana curiosidad erudita, sino de una necesidad apremiante, de un imperativo ético y pragmático. Para regular adecuadamente una tecnología, para anticipar sus riesgos y aprovechar

sus beneficios, para encauzarla hacia el bien común, es menester primero entenderla en toda su complejidad y matices. Y ese entendimiento debe ser profundo, holístico, basado tanto en sus fundamentos teóricos como en su trayectoria histórica.

En las páginas que siguen, nos embarcaremos en esa exploración indispensable, en ese viaje intelectual que es, a la vez, una responsabilidad ineludible. Nos sumergiremos en las raíces conceptuales de la IA, en las ideas seminales de sus pioneros visionarios, en los avances técnicos que han jalonado su evolución, en las fuerzas sociales y económicas que han impulsado su desarrollo. Examinaremos cómo el aumento exponencial de la potencia computacional y la disponibilidad de datos masivos han abierto nuevas fronteras y cómo técnicas clave como las redes neuronales, el aprendizaje automático y el aprendizaje profundo han llevado a la IA a nuevas y asombrosas cotas.

Sección I. En el Principio: los Primeros Pasos de la IA

1.1.1.- Conferencia de Dartmouth y el nacimiento del término “Inteligencia Artificial”

En 1950, el matemático inglés Alan Turing publicó un influyente ensayo titulado *Máquinas de Computación e Inteligencia*, cuyo legado se extendería a la emergente disciplina de la inteligencia artificial. Este suceso tuvo lugar con anterioridad al momento en que la comunidad académica adoptó la expresión "Inteligencia Artificial", acuñada por John McCarthy. El mencionado escrito de Turing se iniciaba con una pregunta aparentemente sencilla: "*¿Pueden las máquinas pensar?*". Seguidamente, el autor sugería un método para evaluar si las máquinas pueden ser dotadas de pensamiento, que se conoció como el Test de Turing. El "*Juego de Imitación*", como fue bautizado en el artículo, fue presentado como una prueba elemental que podría aplicarse para demostrar que las máquinas poseen capacidades mentales. El Test de Turing se enmarca dentro de una perspectiva pragmática, asumiendo que una computadora que resulta indistinguible de un ser humano inteligente ha evidenciado su habilidad para el pensamiento¹⁷. A pesar de que la descripción en cuestión ostenta una aparente simplicidad, las implicaciones derivadas de la construcción de una máquina capaz de superar el Test de Turing son de vasta envergadura. Según

¹⁷ Alan Turing, “*Maquinaria computacional e inteligencia*”, traducido por Cristóbal Fuentes Barassi, Philosophy 36, no. 136 (1950): 433-460, <https://doi.org/10.1017/S0031819100060491>

Turing, se requeriría que esta máquina procesara lenguaje natural, tuviera la capacidad de aprender de la conversación y recordar lo que ha sido expresado, comunicara sus propias ideas al interlocutor humano y comprendiera nociones comunes, evidenciando así lo que se ha denominado sentido común.

Posterior a ello, el punto de inflexión que marcó el inicio formal de la IA como área de estudio fue la icónica Conferencia de Dartmouth, un cónclave intelectual celebrado en el verano de 1956 en el idílico campus del Dartmouth College en Hanover, New Hampshire, Estados Unidos.

Esta conferencia, concebida por el visionario matemático John McCarthy y coorganizada junto a las luminarias Marvin Minsky, Nathaniel Rochester y Claude Shannon, congregó a una pléyade de eminentes científicos y matemáticos con el propósito de "*proceder sobre la base de la conjetura de que cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, ser descrito con tanta precisión que se puede hacer una máquina para simularlo*"¹⁸. La visión de los organizadores era tan audaz como provocadora: alumbrar máquinas capaces de realizar proezas cognitivas que, hasta ese momento, eran coto exclusivo de la mente humana.

Durante las ocho semanas de intensas deliberaciones que abarcó la conferencia, los participantes escrutaron una miríada de temas relacionados con la IA, abarcando desde el procesamiento del lenguaje natural y las redes neuronales hasta la teoría de la computación y la lógica simbólica¹⁹. Aunque los avances tangibles durante el evento en sí fueron más bien modestos, la trascendencia de la Conferencia de Dartmouth radica en haber consagrado la IA como un campo legítimo de investigación y en haber catalizado fecundas colaboraciones entre los pioneros de esta disciplina.

1.1.2- Pioneros y Visionarios: las Primeras Promesas de la IA

Uno de los primeros sistemas de IA que emergieron en esta época germinal fue el *Logic Theorist*, un ingenio computacional alumbrado por el tandem formado por Allen Newell, Herbert

¹⁸ James Moor, "La conferencia de inteligencia artificial del Dartmouth College: los próximos cincuenta años", AI Magazine 27, núm. 4 (2006): 87-91, <https://onlinelibrary.wiley.com/doi/pdf/10.1609/aimag.v27i4.1911>

¹⁹ Ibid, p. 89.

A. Simon²⁰ y Cliff Shaw en 1956²¹. Este programa exhibía la proeza de demostrar teoremas matemáticos a partir de los axiomas contenidos en los *Principia Mathematica* de Alfred North Whitehead y Bertrand Russell. Aunque su desempeño distaba de ser óptimo, el Logic Theorist constituyó un hito al erigirse como el primer sistema en emplear heurísticas y búsqueda de objetivos para desentrañar problemas de envergadura.

Otro jalón destacado en esta época fue el advenimiento del General Problem Solver (GPS), un artefacto computacional concebido por Newell y Simon en 1957²². El GPS fue pergeñado como un sistema de resolución de problemas de propósito general, aplicable a un amplio espectro de dominios, desde la geometría hasta el ajedrez. Su funcionamiento se basaba en una técnica denominada "análisis de medios y fines", que consistía en descomponer un problema en subobjetivos más manejables para luego resolverlos de manera incremental. Este modelo "goal-based" hunde sus raíces en los escritos de Aristóteles sobre la relación entre el conocimiento y la acción²³. En su tratado *De Motu Animalium (Sobre el movimiento de los animales)*, Aristóteles postula que las acciones encuentran su justificación en una conexión lógica entre los fines anhelados y el conocimiento de las consecuencias de cada acto. El filósofo estagirita ilustra esta idea con el siguiente silogismo:

"Pero, ¿cómo ocurre que el pensamiento a veces va acompañado de acción y otras veces no, a veces por movimiento, y otras veces no? Parece como si ocurriese casi lo mismo que en el caso del razonamiento y la realización de inferencias sobre objetos inmutables. Pero en ese caso el fin es una proposición especulativa... mientras que aquí la conclusión que

²⁰ Herbert Simon fue galardonado con el Premio Nobel de Economía en 1978 por introducir el concepto de "satisficing" en la teoría de la toma de decisiones. Este concepto desafió la noción prevaleciente de que los agentes económicos siempre buscan la optimización en sus decisiones. En lugar de perseguir la mejor opción posible mediante un análisis exhaustivo, Simon argumentó que las personas y las organizaciones a menudo se conforman con una solución que es "suficientemente buena", dadas las limitaciones de información, tiempo y recursos cognitivos.

²¹ Allen Newell, J.C. Shaw, y Herbert A. Simon, "Elements of a theory of human problem solving," *Psychological Review* 65, no. 3 (1958): 151-166, https://iiif.library.cmu.edu/file/Simon_box00064_fld04878_bdl0001_doc0001/Simon_box00064_fld04878_bdl0001_doc0001.pdf

²² Allen Newell y Herbert A. Simon, "GPS, a program that simulates human thought," en *Lernende Automaten*, ed. H. Billing (Oldenbourg, 1961), 109-124, https://iiif.library.cmu.edu/file/Simon_box00064_fld04907_bdl0001_doc0001/Simon_box00064_fld04907_bdl0001_doc0001.pdf

²³ Stuart J. Russell y Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3ra ed. (Upper Saddle River, NJ: Pearson Education, Inc., 2010), 22, https://people.engr.tamu.edu/guni/csce421/files/AI_Russell_Norvig.pdf

resulta de las dos premisas es una acción... Necesito abrigo; una capa es un abrigo. Necesito una capa. Lo que necesito, tengo que hacerlo; necesito una capa. Tengo que hacer una capa. Y la conclusión, el "tengo que hacer una capa", es una acción".²⁴

Este silogismo revela cómo el razonamiento práctico parte de una premisa general (la necesidad de una cubierta) y una premisa específica (la identificación del manto como cubierta) para desembocar en una conclusión que es una acción tangible (la confección del manto). Este esquema de razonamiento, en su estructura profunda, prefigura el proceso de descomposición de problemas en subobjetivos y acciones que constituye el núcleo del GPS.

Pero es en la *Ética a Nicómaco* (Libro III.3, 1112b) donde Aristóteles se sumerge en las profundidades de esta cuestión y nos obsequia con un algoritmo de una sutileza y eficacia asombrosas para la deliberación práctica:

"No deliberamos sobre los fines, sino sobre los medios. Pues un médico no delibera si curará, ni un orador si persuadirá... Ellos asumen el fin y consideran cómo y por qué medios se alcanza, y si parece que se produce fácilmente y de la mejor manera por ello; mientras que si se logra por un solo medio consideran cómo se logrará por este y por qué medios se logrará esto, hasta llegar a la primera causa... y lo que es último en el orden del análisis parece ser primero en el orden de la realización. Y si nos encontramos con una imposibilidad, abandonamos la búsqueda, por ejemplo, si necesitamos dinero y esto no se puede obtener; pero si algo parece posible intentamos hacerlo".²⁵

Este pasaje, de una agudeza y anticipación pasmosas, parece una descripción *avant la lettre* del funcionamiento del GPS y otros sistemas de planificación regresiva²⁶. El algoritmo aristotélico

²⁴ Ibid.

²⁵ Ibid.

²⁶ Un sistema de planificación regresiva en inteligencia artificial es un método utilizado para la resolución de problemas y la planificación de tareas. A diferencia de la planificación progresiva, que comienza desde un estado inicial y avanza hacia el estado objetivo a través de una serie de pasos, la planificación regresiva empieza con el objetivo final y trabaja hacia atrás para determinar los pasos necesarios para alcanzar ese objetivo desde el estado inicial. En la planificación regresiva, el proceso comienza con el objetivo o metas finales y se pregunta: "¿Qué acción (o acciones) podría llevarme a este estado objetivo?" Luego, identifica las acciones que podrían resultar en el estado deseado y considera las precondiciones (es decir, los requisitos) para esas acciones. Este proceso se repite hacia atrás, pasando de un conjunto de metas a otro, hasta que se llega a un punto donde las metas se pueden satisfacer directamente por las condiciones en el estado **inicial o mediante acciones simples**.

parte de un fin o estado deseado, y luego escudriña los medios para alcanzarlo, sopesando su viabilidad y optimalidad. Si un medio se revela imposible, se abandona esa senda, pero si parece factible, se intenta materializar. Este proceso se repite, en una recursión de una belleza matemática, hasta arribar a una "causa primera", es decir, una acción directamente ejecutable.

En ese sentido, la arquitectura del GPS se sustentaba en una representación simbólica del conocimiento y en la aplicación de reglas heurísticas para guiar el proceso de búsqueda de soluciones. El sistema partía de un estado inicial y un estado objetivo y mediante la aplicación iterativa de operadores de transformación, exploraba el espacio de posibles soluciones hasta alcanzar el estado deseado.

Newell y Simon argumentaban que este enfoque era análogo a la forma en que los seres humanos resuelven problemas, dividiendo una tarea compleja en subtareas más sencillas y aplicando estrategias generales de resolución. Así, el GPS se erigía como un modelo computacional de la cognición humana, capaz de simular procesos de pensamiento de alto nivel.

Aunque el GPS no logró convertirse en un sistema verdaderamente universal debido a las limitaciones tecnológicas de la época y a la complejidad inherente a ciertos tipos de problemas, su impacto en el desarrollo posterior de la IA fue incommensurable. El GPS sentó las bases conceptuales y metodológicas para el surgimiento de los sistemas expertos y las técnicas de búsqueda heurística. El GPS también tuvo un impacto significativo en la filosofía de la mente y el debate sobre la posibilidad de crear máquinas pensantes. Al demostrar que era posible simular aspectos del pensamiento humano en un sistema computacional, Newell y Simon desafilaron las nociones tradicionales sobre la singularidad de la inteligencia humana y allanaron el camino para la exploración de la IA como un medio para comprender y replicar la cognición.

En el ámbito del procesamiento del lenguaje natural, destaca por su carácter pionero el experimento de Georgetown-IBM de 1954, en el que Leon Dostert y Paul Garvin orquestaron un sistema capaz de traducir automáticamente más de 60 oraciones del ruso al inglés²⁷. Este *tour de force computacional* alumbraba un sistema capaz de traducir automáticamente más de 60

²⁷ J. Hutchins, "The Georgetown-IBM experiment demonstrated in January 1954," en Machine Translation: From Real Users to Research: 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, September 28 – October 2, 2004, eds. Robert E. Frederking y Kathryn B. Taylor (Berlin: Springer Verlag, 2004), 102-114, <https://aclanthology.org/www.mt-archive.info/00/AMTA-2004-Hutchins.pdf>

oraciones del ruso al inglés, una proeza que desafiaba los límites de lo concebible en aquella época primigenia de la informática.

Dostert, un lingüista y traductor de renombre, había sido invitado a la primera conferencia sobre traducción automática en junio de 1952 debido a su vasta experiencia con herramientas mecánicas para la traducción. Su trayectoria era tan impresionante como ecléctica: había fungido como intérprete personal del general Eisenhower durante la Segunda Guerra Mundial, oficiado como enlace con De Gaulle y trabajado para la Oficina de Servicios Estratégicos (predecesora de la Agencia Central de Inteligencia). Tras la guerra, Dostert diseñó e instaló el sistema de interpretación simultánea utilizado en los juicios de Nuremberg y, posteriormente, en las Naciones Unidas, una hazaña técnica y logística que le valió el reconocimiento internacional.

En 1949, la Universidad de Georgetown le extendió a Dostert una invitación para establecer el Instituto de Lenguas y Lingüística en la Escuela de Servicio Exterior, con el mandato de formar una nueva generación de lingüistas y traductores para nutrir las filas del gobierno estadounidense. Fue en este crisol académico donde Dostert, quien había asistido a la conferencia de 1952 como un escéptico, se convirtió en un ferviente apóstol de la traducción automática, convencido de la necesidad imperiosa de explorar sus posibilidades mediante un experimento audaz y práctico.

La elección del par de idiomas para la demostración –del ruso al inglés– obedeció a razones políticas insoslayables en el contexto de la Guerra Fría. La opacidad de las actividades soviéticas era una fuente constante de preocupación para el gobierno estadounidense y Dostert intuía que una demostración exitosa de traducción automática en este par de idiomas estratégicos podría granjearle el apoyo y la financiación necesarios para catapultar el campo a nuevas cotas.

Para materializar su visión, Dostert recurrió a su red de contactos y forjó una alianza improbable con Thomas J. Watson, el visionario fundador de IBM. Juntos, acordaron colaborar en un proyecto que aúna la pericia lingüística de la academia con la potencia computacional de la industria. El timón del proyecto recayó en Cuthbert Hurd, el ingeniero al mando de la División de Ciencias Aplicadas de IBM y en el propio Dostert, quien asumió la dirección científica.

El aspecto lingüístico de este ambicioso experimento quedó en manos de Paul Garvin, un brillante lingüista checo que oficiaba como profesor asociado en el Instituto de Lenguas y Lingüística de Georgetown. Garvin se entregó a la tarea de diseñar un conjunto de oraciones que abarcarán desde la prosa técnica de la química orgánica hasta la expresión cotidiana de temas generales, un corpus cuidadosamente seleccionado para ilustrar una gama de problemas gramaticales y morfológicos que pudieran poner a prueba las capacidades del sistema.

Mientras tanto, en las entrañas de IBM, Peter Sheridan, un programador de talento prodigioso, se afanaba en traducir las reglas lingüísticas de Garvin al lenguaje esotérico de los bits y los bytes. El reto era formidable: conjugar la riqueza y la ambigüedad del lenguaje humano con la rigidez y la precisión de la máquina, todo ello con un vocabulario limitado a 250 elementos léxicos y un puñado de seis reglas gramaticales.

El resultado de esta singular colaboración fue un sistema que, aunque modesto en escala, demostró de manera contundente la viabilidad de la traducción automática. Las más de 60 oraciones traducidas del ruso al inglés, aunque distaban de ser perfectas y requerían una revisión humana extensiva, supusieron un triunfo sin precedentes para la incipiente disciplina de la lingüística computacional y un hito fundacional en la historia de la inteligencia artificial.

El experimento de Georgetown-IBM no solo abrió nuevos horizontes para la investigación en traducción automática, sino que, también, capturó la imaginación del público y encendió el entusiasmo de los círculos gubernamentales y académicos. La prensa se hizo eco de este logro con titulares grandilocuentes que auguraban un futuro en el que las barreras lingüísticas serían derribadas por la férula de la inteligencia artificial; así, el 8 de enero de 1954, la portada del New York Times destacó un reportaje sobre una demostración llevada a cabo el día anterior en la sede de la International Business Machines (IBM) en Nueva York, con el encabezado ***El ruso es convertido al inglés por un avanzado traductor electrónico***:

“Una demostración pública, que se considera la primera implementación exitosa de una máquina capaz de traducir textos con significado de un idioma a otro, tuvo lugar aquí la

*tarde de ayer. Este evento podría representar el punto culminante de siglos de esfuerzos por parte de los académicos en la búsqueda de "un traductor mecánico".*²⁸

Las expectativas iniciales sobre el potencial de la IA eran extraordinariamente altas. Muchos de los pioneros del campo, como Minsky, McCarthy y Simon, predijeron que en pocas décadas se desarrollarían máquinas con una inteligencia comparable o superior a la humana. En 1957, Herbert Simon llegó al extremo de proclamar:

*"No es mi objetivo sorprenderte o impactarte, pero la forma más simple en que puedo resumirlo es decir que ahora existen en el mundo máquinas que piensan, que aprenden y que crean. Además, su capacidad para hacer estas cosas va a aumentar rápidamente hasta que, en un futuro visible, la gama de problemas que pueden manejar será coextensiva con la gama a la que se ha aplicado la mente humana".*²⁹

Posteriormente, en su libro *The Shape of Automation for Men and Management* de 1965, en la misma línea que la declaración anterior, proyectó la siguiente predicción:

*"En un futuro muy cercano, mucho menos de veinticinco años, tendremos la capacidad técnica de sustituir máquinas por cualquier función humana en las organizaciones. En el mismo período, habremos adquirido una teoría extensa y empíricamente probada sobre los procesos cognitivos humanos y su interacción con las emociones, actitudes y valores humanos".*³⁰

Estas proyecciones se materializaron (o casi se materializaron) en un lapso de un poco más de medio siglo, en lugar de las dos décadas inicialmente anticipadas. Estos vaticinios desbordantes de optimismo se cimentaban en los rápidos avances logrados en los primeros años y en una subestimación flagrante de la complejidad inherente a la inteligencia humana. Sin embargo, pronto se hizo patente que la creación de máquinas genuinamente inteligentes entrañaba una dificultad muy superior a la anticipada. Los sistemas de IA desarrollados en las décadas de 1950 y 1960 estaban constreñidos a dominios muy específicos y estructurados y se veían atribulados para lidiar

²⁸ Ibid, p. 102.

²⁹ Stuart J. Russell y Peter Norvig, *Artificial Intelligence: A Modern Approach*, 20.

³⁰ Herbert A. Simon, *The Shape of Automation for Men and Management* (Nueva York: Harper & Row, Publishers, 1965), p, 30, <https://ebin.pub/download/the-shape-of-management-for-men-and-management.html>.

con la ambigüedad, el sentido común y el conocimiento del mundo real que los humanos damos por sentado. Además, la insuficiencia de potencia computacional y la escasez de datos obstaculizaban el desarrollo de sistemas más sofisticados³¹.

A pesar de estos escollos, los primeros años de la IA sentaron un sólido fundamento conceptual y técnico para el desarrollo futuro del campo. Nociones como la búsqueda heurística, la representación del conocimiento, el aprendizaje automático y el procesamiento del lenguaje natural, que hoy constituyen pilares insoslayables de la IA moderna, hunden sus raíces en las investigaciones pioneras de esta época.³²

Además, el entusiasmo inicial y las expectativas elevadas, aunque no se materializaron en el corto plazo, desempeñaron un papel crucial para atraer talento y recursos hacia el campo de la IA. La visión audaz de los fundadores insufló inspiración a generaciones de investigadores para perseguir el sueño de crear máquinas inteligentes, allanando el camino para los prodigiosos avances que hoy presenciamos.

1.2.- El Recorrido Temporal de la IA: una Mirada Retrospectiva a su Evolución

La Inteligencia Artificial (IA) ha experimentado una evolución vertiginosa y apasionante desde sus albores en la década de 1950. Esta trayectoria, jalona por hitos y paradigmas de diversa índole, puede estructurarse en etapas clave que han ido moldeando el campo hasta su fisonomía actual. Cada una de estas fases ha legado conceptos, técnicas y enfoques que han sedimentado en el acervo de la IA, contribuyendo a su paulatina maduración y sofisticación.

1.2.1.- Años 50-60: la Época de la Lógica y las Reglas

Cómo fue reseñado anteriormente, el despegue de la Inteligencia Artificial como disciplina académica suele situarse en la década de 1950, con la célebre conferencia de Dartmouth de 1956 como hito fundacional. En esta primera etapa, que abarca las décadas de 1950 y 1960, la IA estuvo dominada por el paradigma de los **sistemas basados en reglas y la resolución de problemas**

³¹ Russell y Norvig, *Artificial Intelligence: A Modern Approach*, p. 22.

³² Bruce G. Buchanan, "A (very) brief history of artificial intelligence," *AI Magazine* 26, no. 4 (2005): 53-60, <https://doi.org/10.1609/aimag.v26i4.1848>

formales, uno de los primeros enfoques explorados para insuflar capacidades inteligentes en las máquinas.

Estos sistemas codifican el conocimiento en forma de reglas lógicas del tipo "si-entonces", que especifican qué acciones deben acometerse cuando se cumplen determinadas condiciones³³. El motor de inferencia del sistema encadena estas reglas para deducir nuevas conclusiones o tomar decisiones a partir de los datos disponibles.

Para ilustrar este concepto, consideremos un ejemplo del ámbito jurídico. Imaginemos un sistema basado en reglas diseñado para determinar si un individuo califica para recibir asistencia legal gratuita. Una posible regla en este sistema podría expresarse de la siguiente manera:

"Si el ingreso anual del individuo es inferior al umbral de pobreza Y el individuo no posee bienes significativos, ENTONCES el individuo califica para asistencia legal gratuita".

En este caso, la regla establece las condiciones (ingreso anual por debajo del umbral de pobreza y ausencia de bienes significativos) que deben cumplirse para derivar la conclusión (el individuo califica para asistencia legal gratuita). El conjunto completo de reglas en el sistema codificaría el conocimiento experto sobre los criterios de elegibilidad para este beneficio legal.

Algunos de los primeros ejemplos de sistemas de IA basados en reglas fueron los ya expuestos *Logic Theorist (LT)*, alumbrado por el brillante trío formado por Allen Newell, Herbert Simon y Cliff Shaw en 1956 y el *General Problem Solver (GPS)* de Newell y Simon en 1957.

La resolución de problemas formales, como los acertijos lógicos, los juegos estratégicos o las demostraciones matemáticas, fue uno de los principales bancos de pruebas para los sistemas de IA en esta etapa. Se trataba de dominios acotados y bien definidos, donde las reglas y los objetivos podían especificarse de manera precisa. Un ejemplo notable es el programa de A. L. Samuel para el juego de las damas, que era capaz de aprender de la experiencia mejorando su rendimiento³⁴.

³³ Davis, Randall, y Jonathan J. King. "The Origin of Rule-Based Systems in AI." En Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, editado por B. G. Buchanan y E. H. Shortliffe, 20-52. Addison-Wesley, (1984). Recuperado de: <https://www.shortliffe.net/Buchanan-Shortliffe-1984/MYCIN%20Book.htm>

³⁴ Arthur L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," IBM Journal of Research and Development 3, no. 3 (1959): 210-229, <https://doi.org/10.1147/rd.33.0210>

Dicho programa fue un hito en el ámbito del aprendizaje automático y los sistemas de juego. A diferencia de *Logic Theorist (LT)* que operaba en un dominio formal y estático, el programa de Samuel se enfrentaba a un entorno dinámico y adversario, en el que el objetivo era tomar decisiones estratégicas para vencer al oponente.

La innovación clave del programa de Samuel radicaba en su capacidad para aprender y mejorar su rendimiento a través de la experiencia. Utilizando una técnica conocida como aprendizaje por refuerzo, el programa jugaba partidas contra sí mismo y ajustaba sus parámetros de evaluación en función de los resultados obtenidos. Este proceso iterativo de autojuego y ajuste permitía al programa refinar gradualmente su estrategia y mejorar su habilidad para tomar decisiones acertadas³⁵.

El programa de Samuel empleaba una representación del conocimiento basada en funciones de evaluación, que asignaban puntuaciones a diferentes configuraciones del tablero en función de diversas características y patrones. Estas funciones de evaluación eran inicialmente diseñadas a mano, incorporando el conocimiento experto sobre el juego de las damas, pero luego eran refinadas y optimizadas a través del aprendizaje por refuerzo.

Durante el juego, el programa utilizaba estas funciones de evaluación para realizar una búsqueda adelantada, explorando las posibles secuencias de movimientos y seleccionando aquellas que maximizaban su puntuación esperada. Esta búsqueda se realizaba mediante el algoritmo *minimax*³⁶, un enfoque clásico de la teoría de juegos que asume que el oponente siempre seleccionará el movimiento óptimo para minimizar la puntuación del programa.

Más allá de sus logros técnicos, el programa de Samuel tuvo un profundo impacto conceptual en el campo de la IA. Demostró que las máquinas podían aprender y mejorar su rendimiento a través de la experiencia, un principio fundamental del aprendizaje automático que ha impulsado gran parte del progreso posterior en el campo. Además, el enfoque de aprendizaje por refuerzo empleado por Samuel se ha convertido en una técnica estándar en el arsenal de la IA, con aplicaciones que van desde la robótica hasta la toma de decisiones en entornos complejos.

³⁵ Ibid, 218-222.

³⁶ Ibid, 214.

En el ámbito de la resolución de problemas, cabe destacar también el desarrollo de los primeros lenguajes de programación orientados a la IA, como IPL (Information Processing Language) y LISP (List Processing)³⁷, que facilitaron la implementación de algoritmos de búsqueda heurística³⁸ y manipulación simbólica³⁹. Estos lenguajes resultaron especialmente adecuados para expresar las complejas estrategias de razonamiento de los sistemas de IA inicialmente.

Sin embargo, a pesar de los logros iniciales, los sistemas basados en reglas y la resolución de problemas formales mostraron limitaciones importantes. Por un lado, la adquisición y codificación manual del conocimiento resultaba un cuello de botella para el desarrollo de sistemas más amplios y flexibles. Por otro lado, estos enfoques tenían dificultades para manejar la incertidumbre, la ambigüedad y el conocimiento del sentido común necesarios para operar en dominios más realistas.

Además, el entusiasmo inicial y las expectativas desmedidas sobre las capacidades de la IA provocaron una reacción de escepticismo y desencanto cuando los avances no estuvieron a la altura de las promesas. El informe de James Lighthill en 1973⁴⁰ - realizado por solicitud del Comité de Investigación en Ciencia e Ingeniería de Gran Bretaña -, que cuestionaba la viabilidad y la relevancia de la investigación en IA, simbolizó este cambio de percepción y contribuyó a la denominada "*etapa invernal*" (*AI Winter*) que afectó a la disciplina en la década de 1970.

³⁷ John McCarthy, "Recursive Functions of Symbolic Expressions and Their Computation by Machine, Part I," *Communications of the ACM* 3, no. 4 (1960): 184-195, <https://doi.org/10.1145/367177.367199>.

³⁸ La búsqueda heurística se refiere a estrategias de búsqueda que exploran los espacios de soluciones de un problema utilizando métodos que no garantizan encontrar la mejor solución posible pero que, en la práctica, suelen encontrar soluciones buenas en un tiempo razonable. La idea clave detrás de la heurística es utilizar el conocimiento específico del dominio del problema para hacer estimaciones o "mejores conjeturas" sobre cuáles caminos en el espacio de búsqueda podrían llevar más rápidamente a una solución satisfactoria. En el contexto de la IA, los algoritmos de búsqueda heurística son cruciales porque permiten tratar con la vasta complejidad y el tamaño de los espacios de soluciones de muchos problemas. Por ejemplo, en el juego de ajedrez, es impracticable explorar todas las posibles combinaciones de movimientos futuros; en cambio, se utilizan heurísticas para evaluar la posición actual y decidir el mejor movimiento sin necesidad de explorar todas las opciones.

³⁹ La manipulación simbólica es la capacidad de un programa de computadora para trabajar con símbolos o expresiones compuestas por símbolos, en lugar de simplemente con números o valores fijos. Esto permite a las computadoras representar y manipular conceptos abstractos, como los que se encuentran en el lenguaje humano, fórmulas matemáticas o reglas lógicas. Los lenguajes como IPL y LISP fueron diseñados para facilitar la manipulación simbólica, lo que los hizo especialmente adecuados para aplicaciones de IA.

⁴⁰ J. Lighthill, "Artificial Intelligence: A General Survey," en *Artificial Intelligence: A Paper Symposium*, 1-29 (Science Research Council, 1973). https://rodsmithe.nz/wp-content/uploads/Lighthill_1973_Report.pdf

No obstante, a pesar de estas dificultades, la etapa de los sistemas basados en reglas y la resolución de problemas formales estableció muchos de los conceptos, técnicas y enfoques que han seguido siendo fundamentales en el desarrollo posterior de la IA. La noción de representación del conocimiento, los algoritmos de búsqueda heurística, los lenguajes de programación especializados o la importancia del razonamiento simbólico son algunos de los legados duraderos de este período seminal.

1.2.2.- Años 70-80: el Reinado de los Sistemas Expertos

La década de 1970 y 1980 atestiguó el auge de los sistemas expertos y la consolidación de la IA simbólica como paradigma dominante. Los sistemas expertos son

“... un programa de computadora inteligente que utiliza conocimiento e inferencia procedimental para resolver problemas que son lo suficientemente difíciles como para requerir un conocimiento experto humano significativo para su solución. La necesidad de ejecutar a tal nivel, más la inferencia de procedimientos utilizados, puede ser considerada como un modelo de la pericia de los mejores practicantes en el campo. El conocimiento de un sistema experto consiste en hechos y heurísticas. Los "hechos" constituyen un cuerpo de información que es ampliamente compartido, públicamente disponible, y generalmente acordado por expertos en un campo. Las "heurísticas" son principalmente reglas privadas, poco discutidas de buen juicio (reglas de razonamiento plausible, reglas de adivinación buena) que caracterizan la toma de decisiones experta en el campo. El nivel de rendimiento de un sistema experto es primariamente una función del tamaño y la calidad de la base de conocimiento que posee.”⁴¹

⁴¹ Paul Harmon y David King, Expert Systems: Artificial Intelligence in Business (Wiley, 1985), p. 19.
<https://momot.rs/d3/y/1710919988/107/u/annas-archive-ia-2023-06-lcpdf/e/expertsystemsart00harm.pdf~/tl2NkbHLmnQ9TkmimMJsIw/Expert%20systems%3A%20artificial%20intelligence%20in%20business%20--%20Harmon%2C%20Paul%2C%201942-%3B%20King%2C%20David%2C%201949-%20--%204.print.%2C%20New%20York%2C%201985%20--%20New%20York%3A%20J.%20Wiley%20--%209780471808244%20--%2000500ae46cb7987e7b743c0ecbc17cb2b%20--%20Anna%20E%2080%99s%20Archive.pdf>

En este contexto el artículo *The Art of Artificial Intelligence: I. Themes and Case Studies of Knowledge Engineering*⁴², de Edward Feigenbaum se erige como una contribución seminal que captura magistralmente la esencia y el estado del arte de este paradigma en plena efervescencia.

Feigenbaum, con la autoridad que le confiere su rol como pionero y figura señera en el campo⁴³, disecciona con agudeza los principios rectores y las metodologías que constituyen el arte de la ingeniería del conocimiento. Esta disciplina, que emerge como piedra angular de la IA aplicada, se aboca a la tarea de codificar el conocimiento experto en representaciones simbólicas que puedan ser manipuladas por sistemas informáticos para resolver problemas complejos.

El artículo realiza un recorrido comprehensivo y, a la vez, profundamente analítico por los sistemas paradigmáticos gestados en el seno del Heuristic Programming Project de Stanford, auténtico crisol de innovación en IA durante este período. Desde DENDRAL, precursor en la inferencia de estructuras químicas, hasta MYCIN, hito en el diagnóstico médico automatizado, pasando por Meta-DENDRAL, TEIRESIAS y SU/X, Feigenbaum desgrana los principios de diseño, las representaciones de conocimiento y los mecanismos de razonamiento que constituyen la médula de estos sistemas.

Más allá de la descripción pormenorizada de estos hitos, el artículo destila las lecciones cardinales que se derivan de estas experiencias pioneras. El enfoque de generación-y-prueba⁴⁴, las

⁴² E. A. Feigenbaum, *The Art of Artificial Intelligence: I. Themes and Case Studies of Knowledge Engineering* (Stanford, CA: Departamento de Ciencias de la Computación, Universidad de Stanford, agosto de 1977), Memo HPP-77-25, Núm. de reporte STAN-CS-77-621, p. 1-15. <https://stacks.stanford.edu/file/druid:bg342cm2034/bg342cm2034.pdf>

⁴³ Edward Feigenbaum, nacido el 20 de enero de 1936, es una eminencia en el campo de la inteligencia artificial, distinguido por su papel pionero en el desarrollo de sistemas expertos. Obtuvo su doctorado de la Universidad de Carnegie Mellon en 1960, especializándose tempranamente en inteligencia artificial. En la Universidad de Stanford, lideró el proyecto DENDRAL, innovador en la creación de sistemas expertos para la interpretación de espectros de masa en química orgánica. Feigenbaum es reconocido por su teoría de que el conocimiento especializado es esencial para el avance de la IA, planteando que el conocimiento es poder dentro de este ámbito. Esta perspectiva subraya la importancia de la acumulación y aplicación de conocimiento específico en sistemas de IA para resolver problemas complejos. Por sus significativas contribuciones, recibió el Premio Turing en 1994, junto a Raj Reddy, por sus innovaciones en sistemas basados en conocimientos, incluidos los sistemas expertos, que han tenido un profundo impacto tanto académico como industrial.

⁴⁴ El enfoque de generación-y-prueba, también conocido como "*generate and test*", corresponde a una estrategia de resolución de problemas que imita una forma intuitiva de razonamiento humano. Consiste en dos pasos fundamentales: la generación de posibles soluciones (hipótesis) y la prueba de estas soluciones contra un conjunto de criterios para evaluar su viabilidad. En el contexto de la IA, este enfoque permite a los sistemas generar múltiples soluciones potenciales para un problema y luego verificar cada solución hasta encontrar la más adecuada. Este método es especialmente útil en situaciones donde no existe una estrategia directa para resolver un problema, requiriendo en su

reglas de situación-acción⁴⁵ como vehículo para codificar el conocimiento experto y la capacidad de los sistemas para explicar su razonamiento emergen como pilares conceptuales que trascienden los dominios específicos de aplicación.

El desarrollo de estos sistemas que expone Feigenbaum se vio impulsado por varios factores. Por un lado, los avances en la representación del conocimiento, especialmente la introducción de los marcos (frames) y las redes semánticas, permitieron modelar dominios complejos de manera más estructurada y eficiente. Por otro lado, la disponibilidad de hardware más potente y asequible, así como la mejora de las herramientas de programación, facilitaron la implementación de sistemas más ambiciosos.

Los conceptos de “marco” y “redes semánticas” en este ámbito fueron introducidos por Marvin Minsky en 1974 en su artículo *Un marco para representar el conocimiento*⁴⁶. En el contexto de la IA de los años 70, dominada por enfoques basados en lógica y resolución de problemas generales, la propuesta de Minsky de usar estructuras de datos complejas llamadas "marcos" para representar conocimiento supuso un cambio de paradigma. Un marco, según Minsky, es una estructura de datos que representa una situación estereotipada, como estar en un cierto tipo de habitación. Cada marco tiene varios tipos de información adjunta, incluyendo cómo usar el marco, qué esperar que suceda a continuación y qué hacer si estas expectativas no se cumplen.

La idea clave es que un marco no es una simple colección de hechos atómicos, sino una estructura compleja con ranuras (slots) que pueden llenarse con valores por defecto o instancias específicas. Esto permite una representación del conocimiento mucho más rica y flexible que las

lugar una exploración de varias opciones y la selección de la mejor basada en la retroalimentación del proceso de prueba.

⁴⁵ Las reglas de situación-acción - o *production rules* - son un método para codificar el conocimiento experto en sistemas de IA. Este enfoque se basa en la definición de reglas que especifican qué acción debe tomar el sistema en respuesta a una situación o estado particular del entorno o del sistema mismo. Cada regla se formula típicamente como una condición ("si ocurre esta situación") seguida de una acción ("entonces realiza esta acción"). Esta metodología es esencial para la creación de sistemas expertos, ya que permite la representación del conocimiento específico de dominio en una forma que la máquina puede procesar y utilizar para tomar decisiones informadas. Las reglas de situación-acción facilitan la modelización del razonamiento experto, permitiendo a los sistemas responder de manera adaptativa a una amplia gama de escenarios.

⁴⁶ Marvin Minsky, A Framework for Representing Knowledge (MIT-AI Laboratory Memo 306, junio de 1974), reimpresso en The Psychology of Computer Vision, ed. P. Winston (McGraw-Hill, 1975), <https://courses.media.mit.edu/2004spring/mas966/Minsky%201974%20Framework%20for%20knowledge.pdf>

lógicas de predicados típicamente usadas en IA hasta ese momento. Los marcos pueden encapsular conocimiento procedimental sobre cómo usarlos, así como expectativas sobre situaciones típicas, proporcionando una forma poderosa de lidiar con el sentido común y el razonamiento por defecto.

Además, Minsky propone que los marcos relacionados pueden estar vinculados en "*sistemas de marcos*", permitiendo coordinar información desde múltiples perspectivas. Esto introduce la noción de estructurar el conocimiento en redes complejas, en lugar de simples conjuntos de hechos independientes.

El otro concepto clave que Minsky desarrolla son las "*redes de similitud*" o redes semánticas. Propone que los marcos en memoria pueden estar conectados por punteros transformacionales⁴⁷ que representan similitudes o diferencias compartidas entre ellos. Esto forma una densa red de conceptos relacionados por sus propiedades compartidas, permitiendo una organización más eficiente del conocimiento y los procesos como recuperación por similitud o analogía.

Para aclarar la terminología tan técnica que anteriormente se expuso, recurramos a otro ejemplo jurídico: imaginemos que un jurista está tratando un caso de responsabilidad por un accidente automovilístico. Para analizar el caso, el abogado no parte de cero, sino que recurre a un "*marco*" o esquema mental sobre este tipo de situaciones, basado en su experiencia y conocimiento previo.

Este "*marco mental*" del accidente de tráfico incluye varios "*slots*" o ranuras de información esperada, como el tipo de vehículos involucrados, las condiciones de la carretera, las lesiones sufridas, el informe policial, los testimonios de testigos, etc. Algunos de estos slots pueden tener "*valores por defecto*" llenados basándose en situaciones típicas - por ejemplo, normalmente se espera que los conductores implicados tengan seguro.

⁴⁷ Un "puntero transformacional" es un enlace entre conceptos en una red de conocimiento que señala cómo un concepto puede transformarse en otro o cuál es su diferencia principal. Por ejemplo, enlazar "silla" con "taburete" mediante el puntero "sin respaldo" indica que un taburete es como una silla, pero sin respaldo. Estos punteros permiten organizar el conocimiento de manera eficiente, facilitando la recuperación de información por similitud o analogía.

Pero estos valores por defecto pueden ser "*anulados*" por información específica del caso actual. Quizás en este caso uno de los conductores no tenía seguro, lo cual "*rompe*" la expectativa por defecto y requiere un razonamiento adicional.

Además, el marco mental general del "*accidente de tráfico*" está vinculado a otros marcos más específicos en una red de conceptos relacionados. Por ejemplo, puede haber marcos separados para "*colisión frontal*", "*atropello a peatón*", etc., organizados jerárquicamente bajo el marco general de "*accidente*". Cada uno de estos marcos más específicos puede aportar expectativas y razonamientos adicionales al análisis.

Asimismo, el abogado tiene marcos mentales para conceptos como "*negligencia*", "*daños y perjuicios*", "*carga de la prueba*", etc., que están vinculados al marco del accidente por relaciones lógicas. La noción de "*negligencia*", por ejemplo, puede estar vinculada al slot sobre la "*causa*" del accidente.

Durante su análisis, el jurista "*instancia*" o llena estos diversos marcos con la información específica del caso, permitiéndole construir un modelo mental estructurado de la situación. Los slots llenados actúan como "*índices*" para recuperar información relevante, identificar cuestiones clave y razonar por analogía con otros casos.

Adicionalmente, el abogado puede explorar ciertos escenarios o argumentos legales "instanciando" ciertos slots con supuestos hipotéticos y viendo cómo el resto del modelo se actualiza. Esto permite un tipo de razonamiento contrafactual - "*¿Qué pasaría si el conductor hubiera estado ebrio? ¿Cómo cambiaría el análisis?*"

De manera más general, podemos pensar en el corpus completo de conocimiento legal de un jurista (leyes, precedentes, doctrinas, etc.) como una vasta "*red semántica*" de conceptos interconectados. El concepto de "*responsabilidad*", por ejemplo, estaría vinculado a conceptos como "*deber de cuidado*", "*nexo causal*", "*daño*", etc. Cada concepto en la red actúa como un "*nodo*" con "*enlaces*" a conceptos relacionados.

Cuando se enfrentan a un nuevo caso, los abogados trazan un "*camino*" a través de esta red semántica, activando los conceptos y las relaciones más relevantes para construir su argumento.

Los conceptos activados forman un subgrafo que representa la "*región*" de conocimiento legal pertinente para el caso en cuestión.

Ambos conceptos anteriormente expuestos fueron los pilares conceptuales para el desarrollo de los sistemas expertos. Uno de los primeros y más influyentes de esta clase de sistemas fue MYCIN, desarrollado en la Universidad de Stanford a mediados de la década de 1970⁴⁸. MYCIN era un sistema de diagnóstico y recomendación de tratamiento para enfermedades infecciosas de la sangre. Utilizaba una base de conocimiento de alrededor de 450 reglas, adquiridas a partir de entrevistas con expertos médicos, y un motor de inferencia basado en encadenamiento hacia atrás (backward chaining) para razonar sobre los síntomas del paciente y sugerir el tratamiento adecuado.

Un motor de inferencia basado en encadenamiento hacia atrás es un componente de un sistema experto diseñado para razonar o deducir soluciones a partir de una base de conocimientos, empezando por el objetivo o conclusión deseada y trabajando hacia atrás para encontrar los datos o evidencias que soporten esa conclusión.

En el contexto de MYCIN, un sistema experto para el diagnóstico y recomendación de tratamiento de enfermedades infecciosas de la sangre, el encadenamiento hacia atrás permite al sistema comenzar con una hipótesis sobre qué enfermedad o infección puede tener el paciente. Luego, el sistema busca en su base de conocimientos las reglas que podrían llevar a esa conclusión, identificando los síntomas o las condiciones necesarias para que se confirme la hipótesis.

Este proceso implica preguntar al usuario (por ejemplo, un médico o enfermero) por la presencia o ausencia de ciertos síntomas o resultados de pruebas médicas que estén relacionados con la hipótesis en cuestión. El sistema sigue preguntando y razonando hacia atrás hasta que encuentra una cadena de evidencias que soporten (o refuten) la hipótesis inicial, permitiendo así

⁴⁸ Edward Hance Shortliffe, Computer-Based Medical Consultations: MYCIN (New York: American Elsevier Publishing Company, Inc., 1976). Recuperado de: <https://momot.rs/d3/y/1710673429/100/u/annas-archive-ia-2023-06-lcpdf/c/computerbasedmed0000shor.pdf~/HSU8XEUovUQ2oN-z1qecHQ/Computer-based%20medical%20consultations%2C%20MYCIN%20--%20Shortliffe%2C%20Edward%20Hance%20--%201976%20--%20New%20York%3A%20Elsevier%20--%209780444001795%20--%20e266cab310354dd26f341b4e1713fed4%20--%20Anna%20E2%80%99s%20Archive.pdf>

sugerir un diagnóstico y un tratamiento adecuado basado en el conocimiento experto codificado en sus reglas.

Otro hito destacado fue el sistema PROSPECTOR, desarrollado en el Instituto de Investigación de Stanford (SRI) a finales de la década de 1970. PROSPECTOR era un sistema experto para la exploración mineral, capaz de evaluar el potencial de un área para contener depósitos de diversos tipos de minerales⁴⁹.

En el ámbito empresarial, uno de los sistemas expertos más conocidos fue XCON (eXpert CONfigurer), desarrollado por Digital Equipment Corporation (DEC) a principios de la década de 1980⁵⁰. XCON ayudaba en la configuración de los sistemas informáticos VAX⁵¹ de DEC, una tarea compleja que implicaba elegir entre miles de componentes compatibles y satisfacer numerosas restricciones técnicas y de negocio. XCON codificaba el conocimiento de los ingenieros de DEC en forma de reglas y lo aplicaba para generar configuraciones válidas y optimizadas, reduciendo significativamente el tiempo y los errores en el proceso.

Además de los sistemas expertos, la IA simbólica también experimentó avances significativos en otros frentes durante este período. Este paradigma - dominante durante esta década - se fundamenta en la hipótesis de que la cognición humana puede ser modelada y replicada mediante la manipulación de estructuras simbólicas, una idea que encuentra sus raíces en la tradición filosófica del pensamiento simbólico:

“De acuerdo con una tradición fundamental en la filosofía occidental, el acto de pensar (la intelección) consiste esencialmente en la manipulación racional de símbolos mentales,

⁴⁹ A. N. Campbell, V. F. Hollister, R. O. Duda y P. E. Hart, "Recognition of a Hidden Mineral Deposit by an Artificial Intelligence Program," *Science* 217, no. 4563 (1982): 927-929, https://www.jstor.org/stable/1689346?oauth_data=eyJlbWFpbCI6ImtzYW5jaGV6emFtb3JhQGdtYWlsLmNvbSIslmluc3RpdHV0aW9uSWRzIjpBXSwicHJvdmlkZXIiOiJnb29nbGUifQ

⁵⁰ J. McDermott, "R1: A Rule-Based Configurer of Computer Systems" *Artificial Intelligence* 19, no. 1 (1980): 39-88, [main.pdf \(sciedirectassets.com\)](http://main.pdf.sciedirectassets.com)

⁵¹ VAX ("Virtual Address eXtension" o "Extensión de Dirección Virtual") era una línea de superminicomputadoras y sistemas operativos desarrollados por Digital Equipment Corporation (DEC) en los años 70 y 80. Los sistemas VAX fueron conocidos por su potencia, flexibilidad y por el uso del sistema operativo VMS (ahora OpenVMS). Permitían a los usuarios realizar una amplia variedad de tareas computacionales, desde la gestión de bases de datos hasta el desarrollo de software y la simulación científica. Los VAX jugaron un papel crucial en el avance de los sistemas de computación distribuida y en la transición de los sistemas basados en mainframe a arquitecturas de red más descentralizadas.

es decir, de las ideas. No obstante, los relojes (...) no realizan nada que se asemeje en absoluto a la manipulación simbólica racional. Por otro lado, los ordenadores tienen la capacidad de manipular "tokens" arbitrarios de cualquier manera que se especifique; por tanto, parece que solo necesitamos asegurarnos de que esos tokens funcionen como símbolos y que las manipulaciones se definan como racionales para crear una máquina capaz de pensar. En otras palabras, la inteligencia artificial es novedosa y diferente porque los ordenadores efectivamente realizan algo muy similar a lo que se espera que hagan las mentes. De hecho, si esa teoría tradicional es correcta, entonces nuestro ordenador imaginario debería poseer "una mente propia": una mente artificial genuina. ⁵²

Este enfoque concibe la inteligencia como un proceso de razonamiento lógico basado en representaciones declarativas y abstractas del conocimiento y busca desarrollar sistemas que puedan realizar tareas cognitivas de alto nivel, como la resolución de problemas, la planificación y la toma de decisiones, a través de la manipulación explícita de símbolos y reglas.

Durante este período, la IA simbólica experimentó un florecimiento notable, impulsado por los avances en la representación del conocimiento, un área medular que se ocupa de la formalización y estructuración del conocimiento de manera que pueda ser procesado y razonado por sistemas computacionales⁵³. La representación del conocimiento es un desafío fundamental en la IA, ya que determina la capacidad de los sistemas para capturar, organizar y utilizar, de manera efectiva, el conocimiento de un dominio específico.

En este contexto, uno de los desarrollos más significativos en esta época fue la introducción de formalismos lógicos más expresivos y flexibles, que permitieron modelar aspectos del razonamiento y el conocimiento que hasta entonces habían resultado esquivos para los enfoques tradicionales basados en la lógica clásica. Estos formalismos, entre los que destacan las lógicas no monótonas, las lógicas modales y las ontologías, abrieron nuevas vías para representar y razonar

⁵² Haugeland, J. Artificial Intelligence: The Very Idea. MIT Press, (1985), p. 17, https://terragum.com/tfox/books/artificialintelligence_theveryidea.pdf

⁵³ Ronald J. Brachman y Hector J. Levesque, con una contribución de Maurice Pagnucco, Knowledge Representation and Reasoning (San Francisco, CA: Morgan Kaufmann Publishers, 2004), p. 17, <https://www.cin.ufpe.br/~mtcfa/files/in1122/Knowledge%20Representation%20and%20Reasoning.pdf>

con información incompleta, conocimiento incierto, creencias, posibilidades y relaciones conceptuales complejas.

Las lógicas no monótonas⁵⁴, desarrolladas por investigadores como Drew McDermott, Jon Doyle y Raymond Reiter, surgieron como una respuesta a las limitaciones de la lógica clásica para manejar el razonamiento no monótono, es decir, aquel en el que las conclusiones pueden ser retractadas o modificadas a la luz de nueva información. En la lógica clásica, una vez que se infiere una conclusión a partir de un conjunto de premisas, esta se mantiene válida incluso si se agregan nuevas premisas al sistema. Sin embargo, en muchos escenarios del mundo real, el razonamiento es inherentemente no monótono: podemos llegar a conclusiones tentativas basadas en la información disponible, pero estas pueden ser revocadas o ajustadas cuando se obtiene nueva evidencia.

Las lógicas no monótonas introdujeron mecanismos formales para representar y razonar con información incompleta y conocimiento por defecto, permitiendo a los sistemas de IA llegar a conclusiones provisionales en ausencia de información completa y revisar estas conclusiones de manera consistente cuando se agregan nuevos hechos. Estos formalismos, entre los que se encuentran la lógica auto epistémica⁵⁵, la circunscripción⁵⁶ y la lógica por defecto⁵⁷, han encontrado aplicaciones en áreas como el razonamiento de sentido común, la planificación y la representación del conocimiento legal.

Por otra parte, las lógicas modales⁵⁸, cuyo desarrollo fue impulsado por investigadores como Saul Kripke y Jaakko Hintikka, proporcionaron un marco formal para representar y razonar sobre nociones como la necesidad, la posibilidad, el conocimiento y las creencias. A diferencia de la lógica clásica, que se ocupa únicamente de la verdad o falsedad de las proposiciones, las lógicas

⁵⁴ D. McDermott y J. Doyle, "Non-Monotonic Logic I," *Artificial Intelligence* 13, no. 1-2 (1980): 41-72, <https://www.sciencedirect.com/science/article/pii/0004370280900120>

⁵⁵ Ibid, p. 44.

⁵⁶ Ibid, p. 47.

⁵⁷ R. Reiter, "A Logic for Default Reasoning," *Artificial Intelligence* 13, no. 1-2 (1980): 81-132, <https://www.sciencedirect.com/science/article/pii/0004370280900144>

⁵⁸ Ramon Jansana, Lógica Modal (Barcelona: Universitat de Barcelona, n.d.), p. https://www.academia.edu/31570508/L%C3%B3gica_modal

modales introducen operadores adicionales que permiten expresar matices y cualificaciones sobre la modalidad de las afirmaciones:

“Hoy en día lo que se conoce, en sentido amplio, como lógica modal trata de una variedad de modalidades que incluye, además de las tradicionalmente consideradas, otras modalidades que han surgido en las ciencias de la computación y en el estudio de los fundamentos de las matemáticas. Brevemente podemos decir que una modalidad es una expresión que aplicada a una oración S proporciona una nueva oración sobre el modo en que S es verdadera o sobre el modo en que es aceptada. Por ejemplo, sobre cuando es verdadera, donde es verdadera, como es verdadera, en qué circunstancias es verdadera; o sobre el modo en que un sujeto o colectividad la acepta, por ejemplo, como conocida, creida, demostrada, etc.”⁵⁹

Así, por ejemplo, la lógica epistémica permite representar y razonar sobre el conocimiento y las creencias de los agentes, utilizando operadores como "agente A sabe que P" o "agente A cree que P". Esto resulta fundamental en dominios como los sistemas multiagente, donde es necesario modelar y razonar sobre el conocimiento y las creencias de múltiples entidades autónomas que interactúan y se comunican entre sí.

De manera similar, la lógica deóntica, otra variante de la lógica modal, se ocupa de la representación y el razonamiento sobre conceptos normativos como las obligaciones, las permisiones y las prohibiciones:

“La lógica deóntica - a la que es razonable considerar cómo un rama o desarrollo de la lógica modal - se ocuparía de las relaciones de inferencia entre normas, es decir, entre proposiciones prescriptivas”⁶⁰

Esta lógica ha encontrado aplicaciones en el modelado de sistemas legales y éticos, permitiendo formalizar y razonar sobre las normas y consecuencias de las acciones en estos dominios.

⁵⁹ Ibid., p. 5

⁶⁰ Hugo José Francisco Velázquez, "Esclareciendo el concepto de lógica deóntica," Revista Andamios 18, no. 45 (enero-abril 2021): p. 463, https://uacm.edu.mx/portals/5/num45/19_A_Esclareciendo.pdf

Otro avance significativo en la representación del conocimiento durante este período fue el desarrollo de las ontologías⁶¹, impulsado por investigadores como Thomas Gruber y Nicola Guarino. Las ontologías son especificaciones formales y explícitas de una conceptualización compartida, que proporcionan un vocabulario común y una estructura taxonómica para representar el conocimiento de un dominio de manera coherente y reutilizable.

A diferencia de otros formalismos de representación del conocimiento, como las redes semánticas o los marcos, que se centran en las relaciones entre conceptos individuales, las ontologías adoptan una perspectiva más global y sistemática, buscando capturar la estructura conceptual subyacente de un dominio en términos de clases, propiedades, relaciones y axiomas. Esta formalización explícita del conocimiento permite a los sistemas de IA razonar de manera más efectiva sobre las entidades y las relaciones en un dominio, asegurando la consistencia y coherencia de las inferencias realizadas.

Las ontologías han encontrado aplicaciones en una amplia gama de áreas, desde la integración de datos y la recuperación de información, hasta la ingeniería del conocimiento y la web semántica. Al proporcionar un marco compartido para la representación del conocimiento, las ontologías facilitan la interoperabilidad y el intercambio de información entre diferentes sistemas y dominios, un desafío clave en el desarrollo de aplicaciones de IA a gran escala.

Otro hito insoslayable fue el resurgimiento del interés en el procesamiento del lenguaje natural, impulsado por el desarrollo de gramáticas y formalismos lingüísticos más sofisticados. Uno de los sistemas paradigmáticos que encarnó este resurgimiento fue SHRDLU, un prodigo computacional desarrollado por Terry Winograd en el laboratorio de Inteligencia Artificial del MIT a principios de la década de 1970⁶². SHRDLU representó un avance sin parangón en la capacidad de las máquinas para entablar diálogos en lenguaje natural, aunque fuera en el contexto acotado de un microcosmos simplificado conocido como el "*mundo de bloques*".

⁶¹ Thomas R. Gruber, "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition* 5, no. 2 (1993): 199-220, <https://doi.org/10.1006/knac.1993.1008>

⁶² Terry Winograd, "Understanding Natural Language," *Cognitive Psychology* 3, no. 1: 1-191 (Massachusetts Institute of Technology, Cambridge, MA, 1972), <https://www.sciencedirect.com/science/article/pii/0010028572900023?via%3Dihub>

La arquitectura de SHRDLU se sustentaba en un análisis sintáctico y semántico de una finura y complejidad inéditas hasta entonces. Winograd dotó a su sistema de una gramática transformacional aumentada, inspirada en la teoría lingüística de Noam Chomsky, que le permitía analizar la estructura sintáctica de las oraciones con una precisión y una cobertura superiores a las de los enfoques precedentes.

Pero el verdadero prodigo de SHRDLU residía en su capacidad para trascender el mero análisis sintáctico y adentrarse en las sutilezas semánticas del lenguaje. Mediante un intrincado sistema de representación del conocimiento basado en redes semánticas y frames, SHRDLU era capaz de "comprender" el significado de las oraciones en el contexto del mundo de bloques, inferir las implicaciones de las declaraciones y preguntas de los usuarios y generar respuestas coherentes y pertinentes.

Esta hazaña de comprensión lingüística se basaba en una meticulosa codificación del conocimiento sobre los objetos, sus propiedades y las acciones que podían realizarse en el microcosmos de bloques. Cada objeto era representado como un nodo en una red semántica, conectado mediante arcos etiquetados que especificaban sus atributos y relaciones con otros objetos. Los frames, estructuras de datos que agrupaban información sobre situaciones estereotipadas ya expuestas anteriormente, proporcionaban un marco para interpretar las acciones y los eventos descritos en el diálogo.

Gracias a esta rica representación del conocimiento, SHRDLU podía llevar a cabo un razonamiento sofisticado sobre el estado del mundo de bloques y las consecuencias de las acciones realizadas en él. Por ejemplo, si se le pedía que apilara un bloque rojo sobre un bloque verde, SHRDLU podía determinar si esta acción era posible dado el estado actual del mundo, prever los cambios resultantes en la configuración de los bloques y generar una respuesta en lenguaje natural confirmando o explicando la imposibilidad de realizar la acción.

Además de su capacidad para interpretar y responder a declaraciones y preguntas, SHRDLU también podía entablar diálogos interactivos con los usuarios, manteniendo un modelo del contexto conversacional y resolviendo referencias anafóricas. Por ejemplo, si el usuario mencionaba "el bloque rojo" en una oración y luego se refería a "él" en una oración posterior, SHRDLU era capaz de inferir que "él" se refería al bloque rojo mencionado anteriormente.

Esta habilidad para manejar la estructura del discurso y mantener la coherencia en el diálogo fue un logro notable para la época; **SHRDLU demostró que era posible mantener conversaciones en lenguaje natural con una máquina, aunque fuera en un dominio restringido, y que la clave para lograrlo residía en una representación rica y estructurada del conocimiento sobre el mundo.** A pesar de sus impresionantes capacidades, SHRDLU también puso de manifiesto las limitaciones y los desafíos inherentes al procesamiento del lenguaje natural. En primer lugar, el sistema dependía de un conocimiento codificado a mano sobre el mundo de bloques, lo que limitaba su capacidad para escalar a dominios más amplios y complejos. La adquisición y representación del conocimiento de sentido común seguía siendo un cuello de botella formidable para los enfoques simbólicos del PLN.

En segundo lugar, SHRDLU operaba en un entorno artificial y controlado, donde el lenguaje utilizado era relativamente simple y libre de ambigüedades. En contraste, el lenguaje natural en contextos reales está plagado de ambigüedades léxicas y estructurales, expresiones idiomáticas, metáforas y otros fenómenos que desafían el análisis puramente formal. La comprensión profunda del lenguaje requiere no solo un conocimiento lingüístico, sino, también, una comprensión del contexto, la intención comunicativa y el conocimiento del mundo real y el contexto del discurso.

A pesar de estas limitaciones, el impacto de SHRDLU en el campo del procesamiento del lenguaje natural fue profundo y duradero. Demostró el potencial de los enfoques simbólicos basados en gramáticas formales y representaciones estructuradas del conocimiento para modelar aspectos complejos del lenguaje y el razonamiento y allanaron el camino para la investigación posterior en interfaces en lenguaje natural (como el actual y afamado ChatGPT).

Sin embargo, a pesar de los logros de los sistemas expertos y la IA simbólica, esta etapa también puso de manifiesto algunas limitaciones importantes. La adquisición del conocimiento seguía siendo un desafío, ya que extraer y codificar el expertise de los expertos humanos resultaba un proceso arduo y propenso a cuellos de botella. Además, los sistemas expertos tendían a ser frágiles y difíciles de mantener, ya que pequeños cambios en las reglas podían tener efectos cascada difíciles de predecir.

Otra limitación era la dificultad para manejar el conocimiento y razonamiento del “sentido común” necesario para operar en el mundo real. Los sistemas expertos eran muy efectivos en dominios altamente especializados y acotados, pero fallaban al enfrentarse a situaciones novedosas o que requerían una comprensión más amplia del mundo. La obra de Hubert Dreyfus "What Computers Can't Do" expuso de manera influyente estas limitaciones filosóficas de la IA simbólica:

"En lugar de intentar aprovechar las capacidades especiales de las computadoras, los trabajadores en el campo de la inteligencia artificial, cegados por sus primeros éxitos e hipnotizados por la suposición de que el pensamiento es un continuo, no se conformarán con nada menos que con una inteligencia no asistida. La antología de Feigenbaum y Feldman comienza con la declaración más explícita de este principio cuestionable:

"En términos del continuo de inteligencia sugerido por Armer, los programas de computadora que hemos podido construir aún se encuentran en el extremo inferior. Lo importante es que sigamos avanzando en dirección al hito que representa las capacidades de la inteligencia humana. ¿Hay alguna razón para suponer que nunca llegaremos allí? Ninguna en absoluto. No se ha presentado ni una sola pieza de evidencia, ningún argumento lógico, prueba o teorema que demuestre que existe un obstáculo insuperable a lo largo de este continuo".

Armer sugiere con prudencia un límite, pero aun así es optimista:

"Es irrelevante si existe o no un límite superior más allá del cual las máquinas no pueden avanzar en este continuo. Incluso si tal límite existe, no hay evidencia de que esté situado cerca de la posición ocupada por las máquinas actuales".

Las dificultades actuales, una vez interpretadas independientemente de suposiciones optimistas a priori, sugieren, sin embargo, que las áreas de comportamiento inteligente son discontinuas y que el límite está cerca. La estancación en cada uno de los esfuerzos específicos en inteligencia artificial sugiere que no puede haber un avance por partes hacia un comportamiento inteligente adulto plenamente desarrollado para ningún tipo específico de rendimiento humano. Jugar juegos, traducir lenguajes, resolver problemas y

reconocer patrones dependen cada uno de formas específicas de "procesamiento de información" humano, que a su vez se basan en la manera humana de estar en el mundo. Y esta forma de estar-en-una-situación resulta ser, en principio, no programable con las técnicas concebibles actualmente.

Los alquimistas tuvieron tanto éxito destilando mercurio de lo que parecía ser tierra que, después de varios cientos de años de esfuerzos infructuosos para convertir el plomo en oro, aún se negaban a creer que en el nivel químico no se pueden transmutar los metales. Sin embargo, produjeron como subproductos hornos, retortas, crisoles, etc., de la misma manera que los trabajadores informáticos, al no lograr producir inteligencia artificial, han desarrollado programas de ensamblaje, programas de depuración, programas de edición de programas, etc., y el proyecto de robot del M.I.T. ha construido un brazo mecánico muy elegante.

Para evitar el destino de los alquimistas, es hora de preguntarnos dónde estamos parados. Ahora, antes de invertir más tiempo y dinero en el nivel de procesamiento de información, deberíamos preguntar si los protocolos de sujetos humanos y los programas producidos hasta ahora sugieren que el lenguaje informático es adecuado para analizar el comportamiento humano: ¿Es posible un análisis exhaustivo de la razón humana en operaciones regidas por reglas sobre elementos discretos, determinados y libres de contexto? ¿Es siquiera probable una aproximación a este objetivo de la inteligencia artificial? La respuesta a ambas preguntas parece ser, No".⁶³

Además, el entusiasmo por los sistemas expertos también llevó a una cierta sobrevaloración de sus capacidades y a expectativas exageradas sobre su impacto a corto plazo. Cuando muchos proyectos no lograron cumplir con estas expectativas, se produjo una nueva ola de escepticismo y recortes en la financiación de la investigación en IA, conocida como la "segunda etapa invernal" de la IA a finales de la década de 1980 y principios de la de 1990.

⁶³ Hubert L. Dreyfus, What Computers Can't Do: The Limits of Artificial Intelligence (New York: Harper & Row, 1972), p. 214-215,
https://monoskop.org/images/c/ce/Dreyfus_Hubert_What_Computers_Cant_Do_A_Critique_of_Artificial_Reason.pdf

A pesar de los desafíos y limitaciones que enfrentó, la etapa de los sistemas expertos y el auge de la IA simbólica dejó un legado indeleble y transformador en el campo de la inteligencia artificial. Los sistemas expertos, con su capacidad para codificar y aplicar el conocimiento de dominio específico de expertos humanos, demostraron de manera convincente el potencial de la IA para abordar problemas complejos y de alto impacto en el mundo real.

Estos sistemas pioneros, como MYCIN en el diagnóstico médico o PROSPECTOR en la exploración mineral, no solo lograron resultados impresionantes en sus respectivos dominios, sino que, también, sentaron las bases para una nueva forma de concebir la resolución de problemas y la toma de decisiones asistida por computadora. Al capturar y formalizar el conocimiento tácito de los expertos en forma de reglas y estructuras simbólicas, los sistemas expertos allanaron el camino para la creación de repositorios de conocimiento computarizados que podrían preservar y difundir la experiencia humana de manera más amplia y eficiente.

En el ápice de su influencia, la era de los sistemas expertos junto con el florecimiento de la inteligencia artificial (IA) simbólica marcó un parteaguas en la historia de la computación, cimentando su legado a través de su función pionera en la ilustración empírica de la capacidad inherente de la IA para abordar dilemas complejos del mundo real. Este período se distingue no solo por su aportación al avance de metodologías y herramientas vanguardistas para la representación y el razonamiento basados en conocimientos, sino, también, por su visionaria concepción de una inteligencia artificial que pueda capturar, procesar y utilizar el conocimiento humano de manera explícita, estructurada y, sobre todo, efectiva.

La contribución de esta era trasciende la mera acumulación de conocimiento técnico; representa un cambio paradigmático en la percepción del potencial humano y máquina. Al establecer un diálogo entre el intelecto humano y la capacidad computacional, los sistemas expertos y la IA simbólica inauguraron una nueva era de colaboración interdisciplinaria, en la que la filosofía, la lógica, la psicología y la ingeniería convergen en la búsqueda de sistemas autónomos que reflejen la complejidad del razonamiento y la adaptabilidad humana.

1.2.3.- Años 90-2000: el Resurgimiento de las Redes Neuronales

El resurgimiento de la inteligencia artificial, en las postrimerías del siglo XX, estuvo marcado por el redescubrimiento y perfeccionamiento de dos técnicas fundamentales: las redes neuronales artificiales y el aprendizaje automático. Estas herramientas, cuyas semillas conceptuales fueron plantadas décadas atrás por visionarios como Warren McCulloch, Walter Pitts⁶⁴ y Frank Rosenblatt⁶⁵, habían languidecido en el ostracismo durante el invierno de la IA en los 70 y 80. Fue el tenaz trabajo de una nueva generación de investigadores lo que las rescató de su letargo y desencadenó una revolución silenciosa cuyas reverberaciones aún sentimos hoy. Entre estos destacados científicos, cabe mencionar a Geoffrey Hinton y Yann LeCun (actual Científico Jefe de Inteligencia Artificial de Meta, casa matriz de Facebook, Instagram y WhatsApp), cuyas contribuciones seminales allanaron el camino para el renacimiento de estas tecnologías.

Geoffrey Hinton, considerado uno de los padrinos del aprendizaje profundo (Deep Learning) y galardonado con el Premio Nobel de Física en 2024, fue un actor clave en la popularización de las redes neuronales y el algoritmo de retro propagación o "backpropagation"⁶⁶. Sus trabajos pioneros, como el artículo "*Learning representations by back-propagating errors*" publicado en 1986 junto a David Rumelhart y Ronald Williams⁶⁷, sentaron las bases teóricas y prácticas para el entrenamiento eficiente de las redes neuronales.

Por su parte, Yann LeCun es conocido como uno de los pioneros de las redes neuronales convolucionales, una arquitectura especialmente adecuada para el procesamiento de imágenes y la visión por computadora⁶⁸. Sus trabajos seminales, como el desarrollo de la red neuronal LeNet-5

⁶⁴ McCulloch, Warren S., y Walter Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity". *Bulletin of Mathematical Biology* 52, nos. 1/2 (1990): 99-115, <https://www.cs.cmu.edu/~epxing/Class/10715/reading/McCulloch.and.Pitts.pdf>

⁶⁵ Rosenblatt, Frank. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65, no. 6 (1958), [1959-09865-001.pdf \(apa.org\)](https://doi.org/10.1037/h0042518)

⁶⁶ Yann LeCun, Yoshua Bengio, y Geoffrey Hinton, "Deep learning," *Nature* 521, no. 7553 (2015): 436-444, <https://www.nature.com/articles/nature14539>

⁶⁷ David E. Rumelhart, Geoffrey E. Hinton, y Ronald J. Williams, "Learning representations by back-propagating errors," *Nature* 323, no. 6088 (1986): 533-536, <https://www.nature.com/articles/323533a0>

⁶⁸ Yann LeCun, Léon Bottou, Yoshua Bengio, y Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE* 86, no. 11 (1998): 2278-2324, https://axon.cs.byu.edu/~martinez/classes/678/Papers/Convolution_nets.pdf

para el reconocimiento de dígitos manuscritos en la década de 1990⁶⁹, allanaron el camino para los impresionantes avances en tareas como la clasificación de imágenes, la detección de objetos y la segmentación semántica que presenciamos hoy.

Cómo se esbozó al inicio de este apartado, las redes neuronales, experimentaron durante este período un notable perfeccionamiento. En esencia, una red neuronal artificial es un modelo computacional inspirado en la estructura y el funcionamiento del cerebro humano. Consiste en un conjunto de unidades de procesamiento interconectadas, llamadas neuronas artificiales, que colaboran para resolver problemas complejos⁷⁰.

La sofisticación de estas redes neuronales emerge de la interacción sinérgica de sus componentes fundamentales, las neuronas, que actúan como entidades de procesamiento individual dentro de la intrincada estructura de la red. Cada neurona recibe estímulos externos a través de sus conexiones de entrada, los cuales son sometidos a un procesamiento interno que culmina en la generación de un valor de salida.

El mecanismo de computación intrínseco a cada neurona se basa en la realización de una suma ponderada de los valores de entrada. La ponderación de cada conexión de entrada viene determinada por un peso específico, que cuantifica la intensidad con la que cada variable de entrada influye en el comportamiento de la neurona. Así, el peso asignado a cada conexión modula la contribución de la correspondiente variable de entrada en el cómputo global efectuado por la neurona⁷¹.

⁶⁹ Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, y Lawrence D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation* 1, no. 4 (1989): 541-551. <https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf>

⁷⁰ Turmo Borras, Jordi. "Herramientas de extracción de información: redes neuronales". En *Extracción de información en textos escritos en español*, Tesis doctoral, Universidad de Barcelona, (2000), p. 37.. http://deposit.ub.edu/dspace/bitstream/2445/35334/5/3.CAPITULO_2.pdf

⁷¹ "Dot CSV," ¿Qué es una Red Neuronal? Parte 1: La Neurona | DotCSV, YouTube, video, 2:00, 19 de marzo de 2018, <https://www.youtube.com/watch?v=MRIv2IwFTPg&list=PL-Ogd76BhmcB9OjPucsnc2-piEE96jJDQ>

SUMA PONDERADA

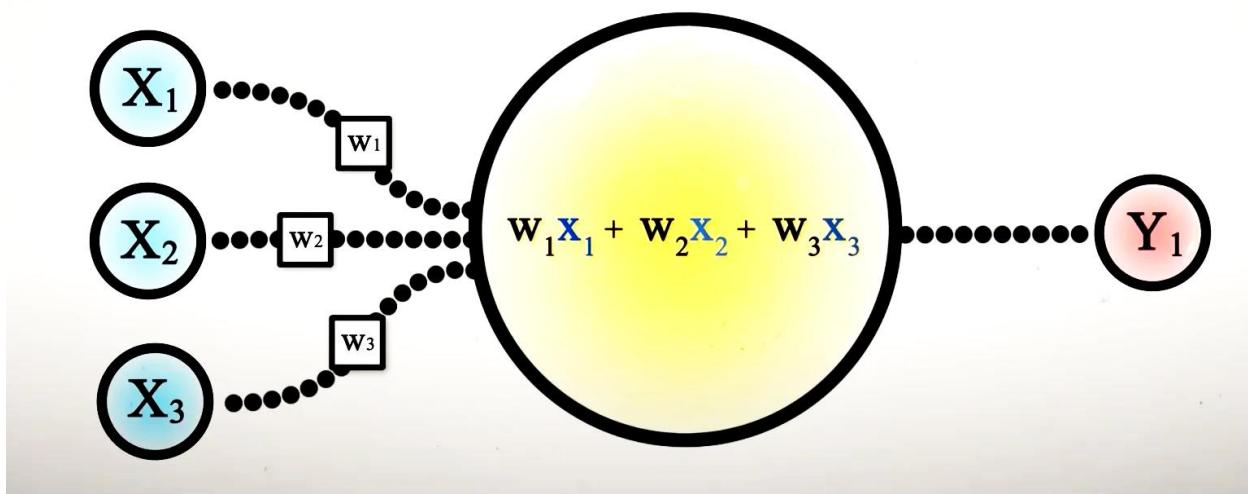


FIGURA 4: Representación esquemática del funcionamiento de una neurona artificial. Las entradas (X_1, X_2, X_3) son ponderadas por sus respectivos pesos (w_1, w_2, w_3) y sumadas. La suma ponderada resultante ($w_1X_1 + w_2X_2 + w_3X_3$) se procesa a través de una función de activación (círculo amarillo) para generar la salida final (Y_1)⁷². Este proceso emula la computación básica realizada por una neurona biológica, donde las entradas son las señales recibidas a través de las dendritas, los pesos representan la fuerza de las sinapsis, y la función de activación determina si la neurona se "dispara" o no, generando una señal de salida.

Para ilustrar de forma más accesible y detallada el funcionamiento de lo anterior, me permitiré esbozar el siguiente ejemplo al lector:

Imagine que tiene tres amigos: Juan, María y Pedro. Cada uno de ellos le da su opinión sobre si debería ir a ver una película en particular. Juan está muy entusiasmado con la película y le da una calificación de 9 sobre 10. María piensa que la película es buena, pero no excepcional, y le da un 7. Pedro, por otro lado, no la recomienda en absoluto y le otorga un 2.

Ahora, usted valora la opinión de cada amigo de manera diferente. Confía mucho en el criterio de Juan para las películas, así que le asigna un peso de 0.5 a su opinión. La opinión de María es importante para usted, pero no tanto como la de Juan, por lo que le da un peso de 0.3.

⁷² Ibid., 2:13.

Finalmente, aunque aprecia a Pedro, rara vez coincide con sus gustos cinematográficos, así que le asigna un peso de 0.2 a su opinión.

Para tomar una decisión, usted realiza una suma ponderada de las opiniones de sus amigos. Multiplica la calificación de cada amigo por el peso que le ha asignado y luego suma estos valores:

$$(9 \times 0.5) + (7 \times 0.3) + (2 \times 0.2) = 4.5 + 2.1 + 0.4 = 7$$

El resultado de esta suma ponderada es 7, que representa la entrada total que usted ha recibido sobre la película.

Ahora, supongamos que tiene un umbral personal para decidir si ve una película o no. Si la entrada total supera un 6, usted va a ver la película. Si es inferior a 6, decide no verla. Esta decisión basada en un umbral se asemeja a la función de activación en una neurona artificial.

En este caso, como la entrada total es 7, que supera su umbral de 6, usted decide ir a ver la película. Esta decisión final (ir o no ir) es análoga a la salida de la neurona artificial.

Volviendo al gráfico, las entradas X1, X2 y X3 representan las opiniones de Juan, María y Pedro, respectivamente. Los pesos w1, w2 y w3 corresponden a la importancia que usted asigna a cada opinión. La suma ponderada ($w_1X_1 + w_2X_2 + w_3X_3$) es el cálculo que usted realiza para combinar las opiniones. Finalmente, la función de activación (el círculo amarillo) representa su umbral personal para tomar una decisión y la salida Y1 es su decisión final de ir o no ir a ver la película.

Este ejemplo simplificado ilustra cómo una neurona artificial procesa información: recibe entradas, las pondera según su importancia, realiza una suma ponderada y luego aplica una función de activación para generar una salida. En una red neuronal, muchas de estas neuronas trabajan juntas, aprendiendo a ajustar los pesos para realizar tareas complejas como el reconocimiento de imágenes o la traducción de idiomas.

Ahora bien, estas neuronas se organizan a lo interno de la red en **capas**, donde las neuronas se disponen en estratos sucesivos. En una red neuronal, las neuronas se disponen en estratos o capas sucesivas, donde cada capa realiza un nivel de procesamiento diferente sobre la información que recibe de la capa anterior.

Típicamente, una red neuronal consta de tres tipos principales de capas⁷³:

- **Capa de entrada:** esta capa recibe los datos o señales externas que se introducen en la red. Cada neurona en la capa de entrada corresponde a una característica o variable de los datos de entrada. Por ejemplo, si la red neuronal está diseñada para reconocer dígitos escritos a mano, cada neurona de entrada podría representar un píxel de la imagen del dígito.
- **Capas ocultas:** son las capas intermedias situadas entre la capa de entrada y la capa de salida. El número de capas ocultas y de neuronas en cada una de ellas puede variar según la complejidad del problema a resolver. En estas capas, las neuronas realizan transformaciones y abstracciones de los datos recibidos de la capa anterior. Cada neurona en una capa oculta recibe entradas ponderadas de las neuronas de la capa previa, aplica una función de activación y envía su salida a las neuronas de la siguiente capa. A medida que la información fluye a través de las capas ocultas, la red va extrayendo características y patrones cada vez más abstractos y complejos de los datos de entrada.
- **Capa de salida:** es la capa final de la red neuronal, donde se produce el resultado o la predicción. El número de neuronas en la capa de salida depende de la tarea que se esté realizando. Por ejemplo, si la red está clasificando imágenes en 10 categorías diferentes, la capa de salida tendría 10 neuronas, cada una correspondiente a una categoría. Las neuronas de la capa de salida reciben entradas ponderadas de las neuronas de la última capa oculta y aplican una función de activación para generar la salida final.

⁷³ "Dot CSV," ¿Qué es una Red Neuronal? Parte 2: La Red | DotCSV, YouTube, video, 1:53, publicado el 26 de marzo de 2018, <https://www.youtube.com/watch?v=uwbHOpp9xkc&list=PL-Ogd76BhmcB9OjPucsnc2-piEE96jJDQ&index=2>

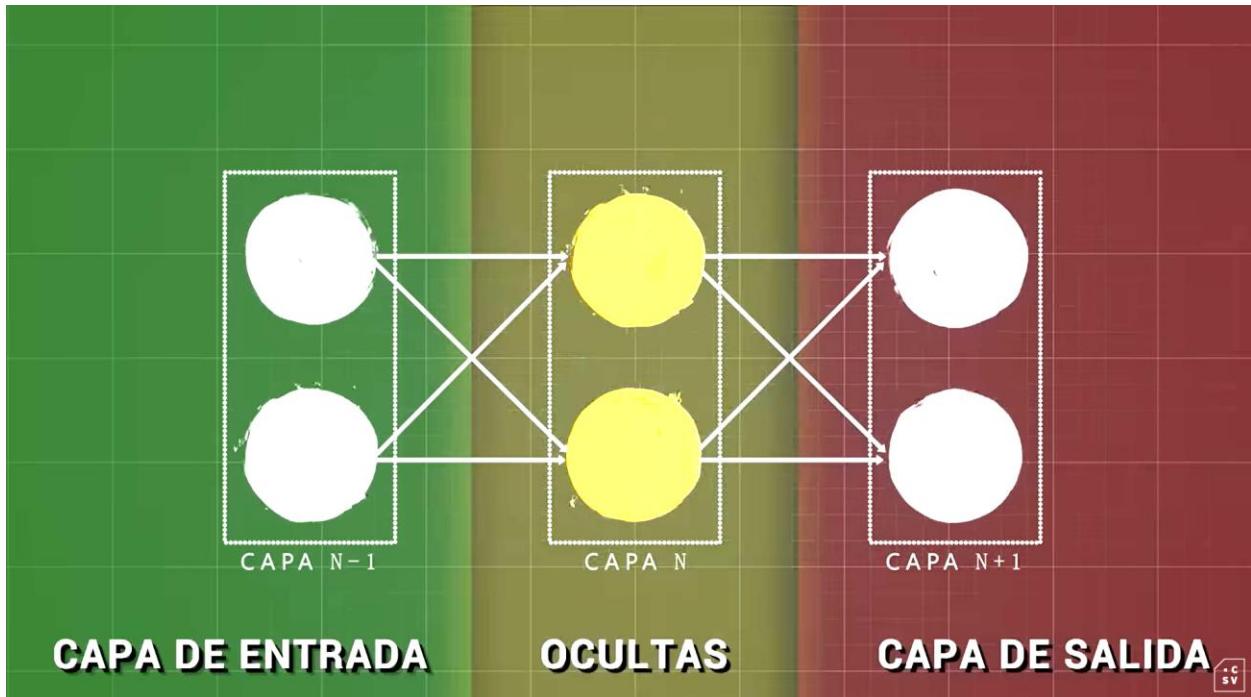


FIGURA 5⁷⁴: Representación esquemática de la arquitectura básica de una red neuronal, destacando la organización en capas. La información fluye desde la capa de entrada (círculos blancos a la izquierda), que recibe los datos sin procesar, hacia la capa oculta (círculos amarillos en el centro), donde las neuronas realizan transformaciones y extraen características más abstractas. Finalmente, la información se transmite a la capa de salida (círculos blancos a la derecha), que genera la predicción o clasificación final. Las flechas representan las conexiones ponderadas entre las neuronas de diferentes capas.

El gráfico presentado ilustra de manera visual la arquitectura básica de una red neuronal, centrándose en la organización en capas que mencionamos anteriormente. En este diagrama, podemos observar tres capas: la capa de entrada, una capa oculta y la capa de salida.

Pensemos que esta red neuronal está diseñada para reconocer frutas en una imagen. La capa de entrada recibiría los píxeles de la imagen como datos de entrada. Cada círculo blanco en la capa de entrada representaría un píxel diferente de la imagen.

A continuación, la información de los píxeles se transmite a la capa oculta, representada por los círculos amarillos. En esta capa, las neuronas realizan transformaciones y extraen características más abstractas de la imagen. Por ejemplo, algunas neuronas podrían especializarse

⁷⁴ Ibid.

en detectar bordes, mientras que otras podrían detectar formas o colores específicos. Cada neurona en la capa oculta recibe información ponderada de todas las neuronas de la capa de entrada y aplica una función de activación para generar su propia salida.

Finalmente, la información procesada en la capa oculta se transmite a la capa de salida, representada por los círculos blancos en el lado derecho. En nuestro ejemplo de reconocimiento de frutas, cada neurona en la capa de salida podría corresponder a una fruta diferente, como una manzana, una banana o una naranja. La red neuronal ajustará los pesos de las conexiones entre las neuronas para que la neurona de salida correspondiente a la fruta correcta se active cuando se le presente una imagen de esa fruta.

De la misma manera en que nosotros adquirimos conocimientos a través de la experiencia y la práctica, refinando nuestras estrategias y métodos para optimizar nuestro rendimiento, las redes neuronales evolucionan mediante la modificación de la intensidad en las conexiones entre sus neuronas. Para clarificar este concepto, utilizaremos un ejemplo adicional. Esto es esencial, ya que esta área de estudio y sus detalles técnicos se encuentran considerablemente alejados de la práctica jurídica.

Figure el lector que está aprendiendo a tocar un instrumento musical, como la guitarra. Al principio, sus movimientos pueden ser torpes y las notas que toca pueden no sonar como desea. Sin embargo, a medida que practica y recibe retroalimentación (ya sea de un instructor o de su propio oído), usted ajusta la forma en que coloca sus dedos, la presión que ejerce sobre las cuerdas y el ritmo con el que toca. Gradualmente, a través de este proceso de ajuste y refinamiento, su habilidad para tocar la guitarra mejora y puede producir melodías cada vez más complejas y armoniosas.

De manera análoga, las redes neuronales aprenden ajustando los pesos de las conexiones entre sus neuronas. Estos pesos determinan la fuerza y la importancia de las conexiones, similar a la fuerza y la precisión con la que usted presiona las cuerdas de la guitarra. Durante el entrenamiento, la red neuronal recibe ejemplos de entrada junto con las salidas deseadas correspondientes. Utilizando algoritmos de aprendizaje, como el de retro propagación, la red compara su salida actual con la salida deseada y calcula la diferencia o el error entre ellas.

A partir de este error, la red realiza ajustes en los pesos de las conexiones, fortaleciendo algunas conexiones y debilitando otras, en un proceso iterativo. Esto es similar a cómo usted ajustaría su técnica de guitarra en función de la retroalimentación que recibe. La red neuronal continúa este proceso de ajuste de pesos, propagando el error hacia atrás a través de las capas de la red y realizando modificaciones en cada iteración, hasta que la salida generada por la red se acerca lo suficiente a la salida deseada.

A través de este proceso de ajuste iterativo de los pesos, la red neuronal va "afinando" sus conexiones, de manera que pueda generalizar y extraer patrones subyacentes en los datos. Así como un guitarrista habilidoso puede tocar canciones que nunca antes ha practicado, una red neuronal bien entrenada puede hacer predicciones precisas o tomar decisiones acertadas sobre datos que nunca ha visto durante el entrenamiento.

El ya citado algoritmo de retro propagación, también conocido como "backpropagation", es el método fundamental que permite el aprendizaje en las redes neuronales artificiales⁷⁵. Este algoritmo engloba todo el proceso de ajuste de pesos y optimización que hemos discutido anteriormente y es el responsable de permitir que las redes neuronales aprendan de manera efectiva a partir de los ejemplos de entrenamiento.

El algoritmo de retro propagación consta de dos fases principales: la propagación hacia adelante (*forward propagation*) y la propagación hacia atrás (*backward propagation*).

1. **Propagación hacia adelante:** en esta primera fase, los datos de entrada, como números o características relevantes para el problema que se está resolviendo, se introducen en la red neuronal. Los datos pasan a través de estas capas, desde la capa de entrada (donde ingresan los datos) hasta la capa de salida (donde se produce el resultado final). En cada capa, las neuronas reciben datos de la capa anterior, los procesan mediante una función matemática (función de activación) y luego envían los resultados a la siguiente capa. Este proceso se repite hasta que se obtiene una salida final.

⁷⁵ "Dot CSV," ¿Qué es una Red Neuronal? Parte 3: Backpropagation | DotCSV, YouTube, video, publicado el 3 de octubre de 2018, https://www.youtube.com/watch?v=eNIqz_noix8&list=PL-Ogd76BhmcB9OjPucsnc2-pIEE96jJDQ&index=4

2. **Propagación hacia atrás:** una vez obtenida la salida final, se compara con la salida deseada o correcta (por ejemplo, en un problema de clasificación, si la red debe identificar si una imagen contiene un gato, la salida deseada sería "gato" o "no gato"). La diferencia entre la salida obtenida y la deseada se llama "error". En esta fase, el objetivo es minimizar este error ajustando los "pesos" de las conexiones entre las neuronas. Los pesos son valores numéricos que determinan la importancia de cada conexión. El proceso de propagación hacia atrás calcula cómo este error se distribuye a través de las conexiones en la red. Utilizando una técnica matemática llamada "regla de la cadena" (o cadena de derivadas), se determina cuánto afecta cada peso al error total. Luego, los pesos se ajustan ligeramente para reducir el error en la próxima iteración.

Este proceso de propagación hacia adelante, cálculo del error y propagación hacia atrás se repite para muchos ejemplos de entrenamiento y las actualizaciones de los pesos se acumulan a lo largo de todo el conjunto de datos. Una pasada completa a través de todo el conjunto de entrenamiento se conoce como una "*época*" (*epoch*)⁷⁶. El algoritmo de retro propagación se ejecuta durante múltiples épocas, y en cada iteración, los pesos se ajustan gradualmente para minimizar el error total en el conjunto de entrenamiento.

A medida que el algoritmo de retro propagación ajusta los pesos de las conexiones, la red neuronal va aprendiendo a mapear las entradas a las salidas deseadas. Los pesos se van adaptando para capturar las relaciones y patrones subyacentes en los datos, permitiendo que la red generalice y haga predicciones precisas incluso en datos no vistos durante el entrenamiento.

Pasemos a otro ejemplo que ilustra el funcionamiento del concepto anterior:

Piense que usted es un profesor que enseña a un grupo de estudiantes a resolver problemas matemáticos. Cada estudiante en la clase representa una neurona en una red neuronal y las conexiones entre los estudiantes representan las conexiones entre las neuronas.

Supongamos que presenta a la clase un problema matemático (los datos de entrada) y les pide que lo resuelvan. Cada estudiante intenta resolver el problema individualmente, utilizando su

⁷⁶ Ibid.

propio enfoque y conocimientos previos (los pesos de las conexiones). Luego, cada estudiante proporciona su respuesta (la salida de cada neurona).

Después de que todos los estudiantes han dado sus respuestas, usted, como profesor, evalúa la precisión de cada respuesta comparándola con la solución correcta (la salida deseada). Si las respuestas no son lo suficientemente precisas, usted les proporciona retroalimentación sobre cómo mejorar.

Aquí es donde entra en juego el algoritmo de retro propagación. En lugar de simplemente decirles a los estudiantes que sus respuestas son incorrectas, usted les da instrucciones específicas sobre cómo ajustar su enfoque para obtener una respuesta más precisa. Por ejemplo, puede sugerirles que presten más atención a ciertos conceptos, que apliquen una fórmula diferente o que revisen sus cálculos.

Los estudiantes toman esta retroalimentación y ajustan su enfoque (actualizan los pesos de las conexiones) en consecuencia. Luego, les presenta un nuevo problema similar y les pide que lo resuelvan nuevamente. Esta vez, los estudiantes incorporan los ajustes que han realizado basados en la retroalimentación anterior.

Este proceso se repite varias veces, con usted proporcionando problemas (datos de entrada) y los estudiantes ajustando continuamente su enfoque basado en la retroalimentación recibida. Con cada iteración, las respuestas de los estudiantes se vuelven más precisas, ya que han aprendido de sus errores anteriores y han ajustado su metodología en consecuencia.

Después de muchas rondas de este proceso, los estudiantes habrán mejorado significativamente su capacidad para resolver este tipo de problemas matemáticos. Han aprendido de los ejemplos y han ajustado su enfoque de manera iterativa hasta que pueden producir consistentemente respuestas precisas.

Este es esencialmente el proceso que sigue el algoritmo de retro propagación en una red neuronal. Los datos de entrada se alimentan a través de la red (los estudiantes intentan resolver el problema) y la salida se compara con la salida deseada (las respuestas se comparan con la solución correcta). El error se calcula y se propaga hacia atrás a través de la red (usted proporciona retroalimentación a los estudiantes) y los pesos de las conexiones se ajustan en consecuencia (los

estudiantes ajustan su enfoque basado en la retroalimentación). Este proceso se repite muchas veces hasta que la red puede producir consistentemente salidas precisas (los estudiantes pueden resolver los problemas correctamente).

Al igual que los estudiantes pueden aprender a resolver nuevos problemas aplicando las habilidades que han adquirido, como se precisó supra, una red neuronal entrenada con el algoritmo de retro propagación puede generalizar su aprendizaje a datos nuevos y no vistos, aprendido a extraer las características y patrones clave de los datos de entrenamiento y puede aplicar este conocimiento a nuevas situaciones.

El algoritmo de retro propagación o "backpropagation" constituye la piedra angular del aprendizaje automático en las redes neuronales, puesto que es el mecanismo que permite a estos modelos computacionales aprender, de manera efectiva, a partir de los ejemplos de entrenamiento. Este algoritmo, cuya elegante simplicidad conceptual contrasta con su profunda trascendencia, ha sido el catalizador que ha impulsado el renacimiento y la explosión de la inteligencia artificial en la última década.

Gracias a este algoritmo, las redes neuronales han demostrado un rendimiento excepcional en una amplia gama de tareas, desde el reconocimiento de imágenes y el procesamiento del lenguaje natural hasta la conducción autónoma de vehículos y el descubrimiento de fármacos. La capacidad de estos modelos para aprender a partir de grandes volúmenes de datos, sin necesidad de una programación explícita, ha abierto un universo de posibilidades y ha impulsado avances sin precedentes en el campo.

Finalizando la época *sub examine*, el partido de ajedrez entre Deepblue y Garry Kasparov, disputado en 1997, marcó un hito histórico en la evolución de esta materia que no podemos ignorar. La victoria de Deepblue sobre el campeón mundial de ajedrez en un torneo oficial fue ampliamente celebrada y generó un gran interés mediático y científico. Este acontecimiento planteó cuestiones cruciales sobre los límites y las posibilidades de la IA en comparación con la inteligencia humana. La segunda vez que se enfrentaron, Deepblue derrotó a Kasparov por un marcador de 3½-2½, lo que se interpretó como un avance simbólico de la IA. El partido también demostró que la creatividad, la intuición y la estrategia de Kasparov seguían siendo una amenaza para la capacidad de la IA de realizar el cálculo exhaustivo de todas las posibles opciones. La capacidad de Deepblue

para analizar hasta 200 millones de posiciones por segundo le proporcionó una ventaja significativa sobre Kasparov en términos de velocidad y precisión. Sin embargo, el primer encuentro demostró que la IA todavía tenía debilidades que podían ser explotadas por un jugador humano.

1.2.4.- 2010-Presente: la Era del Aprendizaje Profundo y los Large Language Models (LLMs)

El último decenio ha presenciado una auténtica revolución copernicana en el dominio de la inteligencia artificial, catapultada por avances disruptivos en el aprendizaje profundo y los modelos de lenguaje de gran escala. Estas innovaciones han trastocado no solo la IA *per se*, sino, también, nuestra comprensión de las capacidades y potencialidades computacionales para emular, e incluso trasvasar, las habilidades cognitivas humanas en un amplio espectro de tareas.

El aprendizaje profundo, un enfoque del aprendizaje automático que emplea redes neuronales artificiales compuestas por múltiples capas ocultas jerárquicamente organizadas (como las previamente expuestas), ha sido el principal impulsor de los vertiginosos avances recientes.

A diferencia de los métodos anteriores que requerían que los humanos identifiquen manualmente las características clave de los datos, el aprendizaje profundo permite que los propios sistemas informáticos descubran, automáticamente, patrones y relaciones complejas directamente a partir de los datos sin procesar. Es como si la máquina pudiera aprender por sí misma en lugar de tener que ser programada explícitamente.

El truco está en utilizar redes neuronales artificiales muy potentes que tienen múltiples capas internas organizadas jerárquicamente. Esto les permite construir su propia comprensión de los datos en niveles cada vez más abstractos y sofisticados.

Cada capa sucesiva extrae patrones cada vez más abstractos y complejos a partir de las representaciones aprendidas en la capa anterior. De este modo, el modelo construye una jerarquía de características que capturan la estructura subyacente de los datos de entrada de manera análoga, aunque exponencialmente más potente, a cómo el cerebro humano procesa la información sensorial.

Esta estructura en capas permite a los modelos de aprendizaje profundo descubrir, automáticamente, las características más discriminativas y relevantes para la tarea en cuestión, sin la necesidad de una ingeniería manual intensiva por parte de expertos humanos. Además, al no depender de una representación predefinida de los datos, estos modelos pueden escalar de manera efectiva para manejar conjuntos de datos masivos y de alta dimensionalidad que superan las capacidades de los enfoques tradicionales de aprendizaje automático⁷⁷.

Un hito fundamental en la aplicación del aprendizaje profundo a la visión por computadora fue la publicación en 2012 del influyente artículo *ImageNet Classification with Deep Convolutional Neural Networks*, por Alex Krizhevsky, Ilya Sutskever (uno de los co-fundadores de la hoy famosa compañía OpenAI) y Geoffrey Hinton⁷⁸. La arquitectura propuesta, conocida como AlexNet, logró una mejora sin precedentes en la tarea de clasificación de imágenes en el conjunto de datos ImageNet, reduciendo el error top-5 del 26 % al 15.3 %. Este logro demostró de forma contundente el inmenso potencial de las redes neuronales profundas para tareas de percepción visual y desencadenó una carrera frenética por aplicar técnicas análogas a una miríada de dominios.

Corolario de lo anterior fue la victoria del sistema AlphaGo (del cual fue co-autor el ya mencionado Ilya Sutskever) de DeepMind - la división de Inteligencia Artificial de Google - en un partido de Go⁷⁹ de cinco juegos contra Lee Sedol, quien generalmente era considerado el mejor jugador del mundo a principios del siglo XXI. Se había anticipado ampliamente que las máquinas eventualmente superarían a los jugadores humanos de Go, como había ocurrido con las computadoras que juegan ajedrez dos décadas antes. Sin embargo, la mayoría de los expertos en Go habían pronosticado que esto no sucedería durante al menos otra década, por lo que el triunfo de AlphaGo fue un momento crucial para el campo de la IA en su conjunto.

⁷⁷ LeCun et al., "Deep Learning."

⁷⁸ Alex Krizhevsky, Ilya Sutskever, y Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in neural information processing systems* 25 (2012): 1097-1105, <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>

⁷⁹ El Go es un juego de mesa estratégico de origen chino que se juega en un tablero de 19x19 líneas, en el que dos jugadores se turnan para colocar piedras negras y blancas en las intersecciones del tablero. El objetivo del juego es controlar una porción mayor del tablero que el oponente, utilizando técnicas de captura y defensa de grupos de piedras.

¿Por qué fue tan relevante esta partida? La respuesta a esta pregunta radica en la complejidad de los problemas que se tienen que resolver en dicho juego. La cantidad de posiciones posibles en el mencionado juego excede con creces el número de átomos que conforman nuestro universo observable, lo que supone una barrera para el análisis exhaustivo de todas las secuencias de movimientos interesantes en un futuro próximo. De tal modo, los jugadores confían en gran medida en la intuición subconsciente para complementar su razonamiento consciente, siendo los expertos capaces de desarrollar una habilidad casi sobrenatural en la identificación de posiciones ventajosas frente a aquellas menos favorables⁸⁰.

Alpha Go logró una simbiosis entre la intuición y la lógica que dio lugar a jugadas que no solo eran poderosas, sino, también, altamente creativas en algunos casos. Por ejemplo, la sabiduría milenaria del Go establece que, al inicio del juego, lo mejor es jugar en la tercera o cuarta línea desde el borde. Hay un compromiso entre ambas líneas: jugar en la tercera línea ayuda a ganar territorio a corto plazo hacia el lado del tablero, mientras que jugar en la cuarta línea ayuda a tener influencia estratégica a largo plazo hacia el centro. En el trigésimo séptimo movimiento del segundo juego, AlphaGo sorprendió al mundo del Go desafiando esa sabiduría antigua y jugando en la quinta línea, como si estuviera aún más seguro que un humano en sus habilidades de planificación a largo plazo y, por lo tanto, favorecía la ventaja estratégica sobre la ganancia a corto plazo. Los comentaristas quedaron atónitos, y Lee Sedol incluso se levantó y salió temporalmente de la habitación. Efectivamente, alrededor de cincuenta movimientos después, la lucha desde la esquina inferior izquierda del tablero terminó desbordándose y conectando con esa piedra negra del movimiento treinta y siete. Y ese motivo fue lo que finalmente ganó el juego, afianzando el legado del movimiento de quinta fila de AlphaGo como uno de los más creativos en la historia del Go⁸¹.

El juego de Go, en virtud de sus aspectos intuitivos y creativos, se reconoce en Oriente como una forma de arte más que meramente un juego. Históricamente, en la China antigua, se consideró una de las cuatro artes esenciales, junto con la pintura, la caligrafía y la música qin, y a la fecha, sigue siendo ampliamente apreciado en Asia, con un aproximado de 300 millones de espectadores durante el primer encuentro entre AlphaGo y Lee Sedol. Por lo tanto, el mundo del

⁸⁰ Max Tegmark. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Alfred A. Knopf, 2017, p. 114. Recuperado de: [Life 3.0: Being Human in the Age of Artificial Intelligence \(readyforai.com\)](http://Life 3.0: Being Human in the Age of Artificial Intelligence (readyforai.com))

⁸¹ Ibid, p. 116.

Go quedó sumamente agitado por el resultado del encuentro y consideró la victoria de AlphaGo como un hito trascendental para la humanidad. El jugador de Go más importante de ese momento, Ke Jie, expresó que

"La humanidad ha practicado el Go durante miles de años y, sin embargo, como nos ha demostrado la IA, aún no hemos Arañado la superficie... La unión de jugadores humanos y computadoras marcará el comienzo de una nueva era... Juntos, el hombre y la IA pueden encontrar la verdad del Go"⁸².

En el campo del procesamiento del lenguaje natural, el advenimiento de una nueva arquitectura de aprendizaje profundo supuso un cambio de paradigma en la materia,

Los “Transformers” fueron presentados por primera vez en un artículo llamado *“Attention Is All You Need”*⁸³ en 2017, y desde entonces se han convertido en la técnica predominante para muchas tareas de NLP (*Natural Language Processing*), como la traducción automática, el resumen de texto y el análisis de sentimientos. La clave de los Transformers es su capacidad para aprender patrones en el texto utilizando atención, una técnica que permite a la red neural enfocar su atención en diferentes partes del texto y aprender relaciones entre ellas. Esto significa que los Transformers pueden procesar información a nivel de oración o de párrafo completo, en lugar de analizar las palabras una por una, lo que hace que el procesamiento sea mucho más eficiente. Además, los Transformers también han demostrado ser muy efectivos para aprender representaciones vectoriales⁸⁴ de palabras, también conocidas como “embeddings”. Estos embeddings son altamente informativos y útiles para una variedad de tareas de NLP. Cuando entrenamos un modelo Transformer con grandes conjuntos de datos, el modelo aprende a convertir las palabras en vectores numéricos. Estos vectores capturan tanto el significado semántico (lo que las palabras significan) como el sintáctico (cómo se utilizan las palabras en una oración) de las palabras. Esta

⁸² Ibid, p. 117.

⁸³ Vaswani, Ashish, et al. "Attention Is All You Need." Advances in Neural Information Processing Systems, 2017, Recuperado de: [\[1706.03762\] Attention Is All You Need \(arxiv.org\)](https://arxiv.org/abs/1706.03762)

⁸⁴ Las representaciones vectoriales son una forma de representar información en un formato matemático, específicamente en forma de vectores numéricos. En el contexto del procesamiento del lenguaje natural (NLP), las representaciones vectoriales se utilizan para representar el significado de las palabras o frases en un espacio matemático continuo.

capacidad de comprender el contexto y el significado de las palabras permite que los modelos Transformer realicen tareas complejas de NLP con mayor precisión⁸⁵.

En este proceso cada palabra en el vocabulario del modelo se representa mediante un vector, que es esencialmente una lista de números. Estos números no son aleatorios; están ajustados de tal manera que capturan tanto el significado de la palabra como el contexto en el que suele aparecer. Por ejemplo, palabras como "gato" y "perro" tendrán vectores que son más cercanos entre sí en comparación con palabras no relacionadas como "mesa" o "cielo". Esto se debe a que, durante el entrenamiento, el modelo aprende a ajustar los valores de los vectores para reflejar relaciones semánticas y sintácticas.

Una de las grandes ventajas de los Transformers es su capacidad para procesar palabras en contexto. No solo consideran las palabras de manera aislada, sino que, también, tienen en cuenta su relación con las demás palabras en una oración. Esto significa que pueden distinguir entre diferentes significados de una misma palabra dependiendo de cómo se usa en una oración. Por ejemplo, la palabra "banco" tendrá diferentes vectores si se usa en el contexto de "institución financiera" o "asiento en un parque".

Gracias a estos embeddings de alta calidad, los modelos Transformer pueden realizar tareas complejas de NLP con una precisión mucho mayor. En la clasificación de texto, por ejemplo, pueden analizar los vectores de las palabras en un documento para determinar su categoría, como si un artículo es sobre política, deportes o tecnología. En el análisis de sentimientos, pueden identificar si el tono de un texto es positivo, negativo o neutral, basándose en las relaciones entre las palabras y su contexto.

Esta arquitectura, además de lograr un rendimiento superior en tareas de traducción automática, demostró una escalabilidad y eficiencia computacional sin precedentes, allanando el camino para el surgimiento de los Large Language Models o Modelos de Lenguaje de Gran Tamaño.

⁸⁵ Ibid.

2.- Los Large Language Models: Reescribiendo las Reglas del Juego

Los LLMs, representan la vanguardia actual en el modelado del lenguaje natural. Estos titanes computacionales, con cientos de miles de millones de parámetros entrenados en vastos corpus de texto, han demostrado una capacidad asombrosa para generación de lenguaje coherente, comprensión de contexto y razonamiento analógico. GPT-3, en particular, ha cautivado la imaginación del público con su habilidad para realizar una amplia gama de tareas lingüísticas, desde la traducción y el resumen hasta la generación de código y la escritura creativa, todo a partir de instrucciones en lenguaje natural y con una supervisión mínima⁸⁶.

2.1.- GPT 3, GPT 3.5 y GPT-4: Buques Insignia de OpenAI

El advenimiento de los LLMs puede trazarse hasta el lanzamiento de GPT-3 (Generative Pre-trained Transformer 3) por OpenAI en 2020, un modelo que marcó un antes y un después en el campo. Con sus 175 mil millones de parámetros y su entrenamiento en un vasto corpus de texto no estructurado, GPT-3 fue desarrollado por OpenAI con el objetivo inicial de impulsar avances en la generación de lenguaje natural, la traducción automática y las tareas de autocompletado⁸⁷. Sin embargo, a medida que el modelo fue desplegado y estudiado más a fondo, se hizo evidente que sus capacidades se extendían mucho más allá de estos objetivos iniciales, abarcando una amplia gama de habilidades emergentes que sorprendieron incluso a sus propios creadores.

El entrenamiento de GPT-3 siguió un enfoque en el que el modelo es entrenado para predecir la siguiente palabra en una secuencia dado el contexto anterior. Este enfoque simple pero poderoso permite al modelo aprender patrones y relaciones profundas en el lenguaje de una manera no supervisada, sin la necesidad de un etiquetado manual extensivo⁸⁸. Como resultado, GPT-3 adquirió una comprensión robusta de la sintaxis, la semántica y el contexto del lenguaje, que forma la base de sus impresionantes capacidades de generación y comprensión del lenguaje.

⁸⁶ Alex Tamkin, Miles Brundage, Jack Clark, y Deep Ganguli, "Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models," arXiv:2102.02503v1 [cs.CL] (4 de febrero de 2021), [2102.02503.pdf \(arxiv.org\)](https://arxiv.org/pdf/2102.02503.pdf)

⁸⁷ Tom B. Brown, et al., "Language models are few-shot learners," arXiv preprint arXiv:2005.14165 (2020), <https://arxiv.org/abs/2005.14165>

⁸⁸ Alec Radford, Karthik Narasimhan, Tim Salimans e Ilya Sutskever, "Improving Language Understanding by Generative Pre-Training" (OpenAI, 2018), <https://www.cs.ubc.ca/~amuhamed/LING530/papers/radford2018improving.pdf>

Inicialmente, GPT-3 fue aplicado a tareas como la traducción automática y el autocompletado, donde demostró un rendimiento excepcional. En la traducción automática, GPT-3 pudo generar traducciones de alta calidad para una amplia gama de pares de idiomas, a menudo superando a los sistemas de última generación específicos para la traducción⁸⁹. En el autocompletado, GPT-3 mostró una notable capacidad para generar continuaciones de texto coherentes y contextualmente apropiadas, con aplicaciones potenciales en la escritura asistida, la generación de contenido y más⁹⁰.

Sin embargo, a medida que los investigadores y usuarios comenzaron a experimentar con GPT-3, se hizo evidente que sus capacidades se extendían mucho más allá de estas tareas iniciales. Se descubrió que GPT-3 podía realizar una amplia gama de tareas de procesamiento del lenguaje natural, a menudo con pocos o ningún ejemplo, un fenómeno conocido como aprendizaje de pocos disparos (*Few-Shot-Learning*)⁹¹. Esto incluía tareas como la respuesta a preguntas, el resumen de texto, la inferencia de sentimientos, e incluso la generación de código y la resolución de problemas matemáticos simples.

Además, GPT-3 demostró capacidades sorprendentes en tareas que van más allá del lenguaje, como el razonamiento relacional, la planificación y la creatividad⁹². Por ejemplo, dado un conjunto de relaciones textuales, GPT-3 puede razonar sobre las relaciones implícitas y sacar conclusiones, una habilidad que se pensaba que requería una representación simbólica explícita. En tareas de planificación, GPT-3 puede generar secuencias de acciones plausibles para lograr un objetivo dado, demostrando una comprensión de la causalidad y la temporalidad. Y en tareas creativas, GPT-3 mostró una capacidad notable para la generación de ideas, historias y conceptos novedosos.

⁸⁹ Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, y Hany Hassan Awadalla, "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation," Microsoft, febrero de 2023, DOI: 10.48550/arXiv.2302.09210, https://www.researchgate.net/publication/368664574_How_Good_Are_GPT_Models_at_Machine_Translation_A_Comprehensive_Evaluation/fulltext/63f4374f57495059452fbe19/How-Good-Are-GPT-Models-at-Machine-Translation-A-Comprehensive-Evaluation.pdf?tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19

⁹⁰ Gwern Branwen, "GPT-3 Creative Fiction," 2020, <https://www.gwern.net/GPT-3>

⁹¹ Brown et al., 2020.

⁹² Luciano Floridi y Massimo Chiriaci, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds and Machines* 30, no. 681 (2020): 694, <https://doi.org/10.1007/s11023-020-09548-1>

Estas capacidades emergentes tomaron por sorpresa incluso a los creadores de GPT-3, quienes reconocieron que el modelo estaba exhibiendo un comportamiento que iba más allá de lo que se pretendía o anticipaba⁹³. Esto llevó a un intenso debate y especulación sobre la naturaleza de la inteligencia exhibida por GPT-3 y otros grandes modelos de lenguaje. Algunos argumentaban que estos modelos están exhibiendo una forma de comprensión y razonamiento genuinos, mientras que otros sostienen que simplemente están aprovechando patrones estadísticos complejos en los datos.⁹⁴ A pesar de su impresionante rendimiento, el modelo a menudo lucía mal en sus respuestas, divagaba sin un claro sentido de objetivo, y a veces generaba texto que era factualmente incorrecto, inconsistente o sesgado. Estas limitaciones subrayaron la necesidad de mejorar la robustez, la coherencia y la alineación de valores de los LLMs, un desafío que OpenAI abordó de frente con el desarrollo de GPT-4, pero antes de eso estuvo ChatGPT.

ChatGPT⁹⁵, basado en la arquitectura GPT-3.5, marcó un hito trascendental en la evolución de los modelos de lenguaje y su interacción con el público general. Lanzado por OpenAI el 30 de noviembre de 2022, rápidamente capturó la atención masiva por su capacidad para entablar conversaciones fluidas, coherentes y aparentemente naturales con los usuarios, todo a través de una intuitiva interfaz de chat.

Bajo el capó, ChatGPT se sustenta en una versión optimizada de GPT-3.5, una iteración de la arquitectura GPT-3 que había sido objeto de un intenso refinamiento y mejora desde su lanzamiento inicial. A diferencia de GPT-3, que fue entrenado principalmente mediante aprendizaje supervisado en un vasto corpus de texto de Internet, GPT-3.5 incorporó técnicas adicionales de entrenamiento, como el aprendizaje por refuerzo con retroalimentación humana (**RLHF**, por sus siglas en inglés).

En esencia, RLHF implica ajustar el modelo no solo para predecir el siguiente token en una secuencia, sino para generar respuestas que se alineen con las preferencias y los valores

⁹³ Rylan Schaeffer, Brando Miranda, y Sanmi Koyejo, "Are Emergent Abilities of Large Language Models a Mirage?" (Computer Science, Stanford University, mayo de 2023), arXiv, [2304.15004.pdf \(arxiv.org\)](https://arxiv.org/pdf/2304.15004.pdf)

⁹⁴ Gary Marcus y Ernest Davis, "GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About," (MIT Technology Review, 22 de agosto de 2020), <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>

⁹⁵ OpenAI, "Introducing ChatGPT," blog de OpenAI, 30 de noviembre de 2022, accedido el 18 de mayo de 2024, <https://www.openai.com/blog/chatgpt>

humanos. Esto se logra mediante un proceso iterativo en el que el modelo genera múltiples respuestas candidatas, que luego son clasificadas por anotadores humanos según su calidad y adecuación. Estas clasificaciones se utilizan para calcular una "recompensa" que guía al modelo hacia la generación de respuestas más deseables⁹⁶.

Para recopilar los datos de comparación necesarios para este enfoque, OpenAI recurrió a una configuración ingeniosa. Los instructores humanos entablaron conversaciones con el chatbot, asumiendo tanto el papel del usuario como del asistente de IA. Los instructores tuvieron acceso a sugerencias generadas por el modelo para ayudarlos a componer sus respuestas. Luego, OpenAI seleccionó aleatoriamente un mensaje escrito por el modelo, muestreó varias alternativas de completado y pidió a los instructores que las clasificaran. Estos rankings sirvieron como base para los modelos de recompensa utilizados en el ajuste fino del modelo.

Además de RLHF, GPT-3.5 también se benefició del ajuste fino supervisado en un conjunto de datos de diálogo especialmente diseñado. Los instructores humanos proporcionaron conversaciones en las que desempeñaron ambos roles, creando ejemplos de interacciones ideales entre usuarios y asistentes de IA. Este conjunto de datos de diálogo se combinó con el conjunto de datos InstructGPT, transformado a un formato de diálogo, para crear un recurso de entrenamiento completo y diverso.

El resultado de este proceso de entrenamiento multifacético fue un modelo que no solo podía generar respuestas de alta calidad, sino que, también, exhibía un notable grado de adaptabilidad y alineación con las intenciones y preferencias del usuario. ChatGPT demostró habilidad para entender el contexto, mantener la coherencia a lo largo de interacciones prolongadas y ajustar su tono y estilo para adaptarse a las necesidades de la conversación.

Pero quizás el aspecto más innovador de ChatGPT fue la forma en que democratizó el acceso a la potencia de los LLMs. Al encapsular las capacidades de GPT-3.5 en una interfaz de chat accesible y fácil de usar, OpenAI puso esta tecnología transformadora al alcance de millones de usuarios no técnicos. De repente, cualquier persona con una conexión a Internet podía

⁹⁶ Ibid.

interactuar con un sistema de IA de vanguardia, explorando sus vastas capacidades y obteniendo conocimientos fascinantes en el proceso.

La interfaz de ChatGPT fue diseñada para ser intuitiva y atractiva, con un minimalismo que ponía el foco en la conversación. Los usuarios simplemente escribían sus mensajes en un cuadro de chat y ChatGPT respondía en tiempo real con respuestas bien formadas y sustanciales. El sistema podía manejar una amplia gama de consultas, desde preguntas fácticas hasta tareas de escritura creativa, pasando por resolución de problemas y análisis.

Otra de las características más notables de ChatGPT era su capacidad para mantener un hilo coherente a lo largo de interacciones prolongadas. A diferencia de los chatbots tradicionales, que a menudo luchaban por mantener el contexto más allá de unos pocos intercambios, ChatGPT podía seguir una línea de discusión, hacer referencias a puntos planteados anteriormente y construir sobre ideas de una manera que se sentía asombrosamente natural y conversacional.

Pero GPT-3.5 y ChatGPT no fueron sino el preludio de una nueva era en el modelado del lenguaje. Lanzado en 2023, GPT-4 representa un salto cuántico en las capacidades de los LLMs⁹⁷. Con una arquitectura refinada y una escala aún mayor, GPT-4 no solo supera a su predecesor en una amplia gama de tareas y benchmarks⁹⁸, sino que, también, exhibe una comprensión más profunda y matizada del lenguaje y el contexto. Las mejoras en GPT-4 son tanto cuantitativas como cualitativas: el modelo no solo comete menos errores y genera texto más coherente, sino que, también, demuestra una capacidad impresionante para el razonamiento analógico, la resolución de problemas complejos y la generación de insights originales.

De manera notable, y especialmente relevante para el tema que nos ocupa, GPT-4 ha superado una versión simulada del examen de admisión a la práctica del derecho en los Estados Unidos (Uniform Bar Examination) con un puntaje que lo sitúa en el 10% superior de los humanos que han tomado esta prueba.

La trascendencia del éxito alcanzado por GPT-4 en el examen de admisión a la práctica del derecho radica en el espectacular nivel de precisión y desempeño demostrado por este modelo, lo

⁹⁷ OpenAI, “GPT-4 Technical Report”, (Marzo de 2023). Recuperado de: <https://arxiv.org/abs/2303.08774>.

⁹⁸ El término "benchmark" se refiere a una medida de referencia que se utiliza para comparar el rendimiento de un sistema, producto o servicio con respecto a otros similares en el mercado o a un estándar establecido.

cual, a su vez, augura un horizonte prometedor en la aplicación de sistemas de procesamiento de lenguaje natural en el ámbito jurídico. Estas posibilidades, que hasta hace poco parecían obstaculizadas por dificultades técnicas insuperables, se ven ahora fortalecidas por el extraordinario avance logrado por este modelo.

De hecho, al comparar su desempeño en la misma prueba con su modelo predecesor, GPT-3.5, se constata una tendencia al aceleramiento de la evolución de estos sistemas, que se erigen como herramientas valiosas para la automatización de procesos en el campo del derecho y la optimización de la eficiencia y precisión en la toma de decisiones .

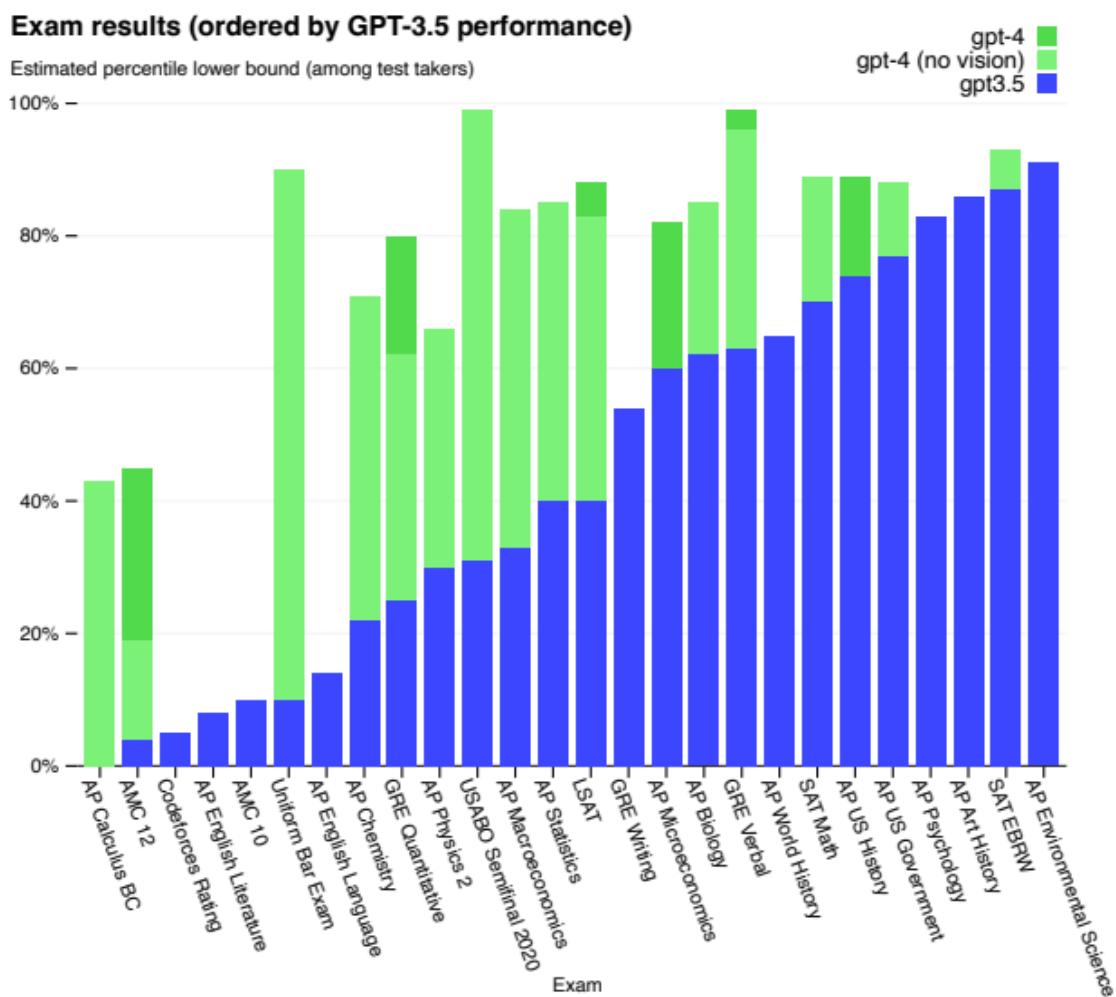


FIGURA 6: Desempeño de GPT en exámenes académicos y profesionales. En cada caso, se simularon las condiciones de evaluación y puntuación del examen real. Los exámenes se ordenan de menor a mayor según el desempeño de GPT-3.5. GPT-4 supera a GPT-3.5 en la mayoría de los exámenes evaluados⁹⁹.

Otra de las innovaciones clave de GPT-4 es su capacidad para procesar y generar no solo texto, sino también imágenes y otros tipos de datos multimedia. Esta habilidad multimodal abre un abanico de nuevas aplicaciones, desde la generación de descripciones de imágenes para personas con discapacidad visual hasta la creación de contenido multimedia rico y atractivo. Sin embargo, a pesar de estos avances, GPT-4 también ha suscitado preocupaciones y debates. Algunos expertos han advertido sobre los riesgos potenciales de un modelo tan poderoso y autónomo, desde la generación de desinformación a gran escala hasta la automatización del trabajo cognitivo¹⁰⁰.

2.2.- Claude 2 de Anthropic y las Ventanas de Contexto

Pero OpenAI no es el único actor en este escenario que evoluciona exponencialmente. La salida al mercado de Claude 2, la penúltima iteración del modelo de lenguaje desarrollado por Anthropic (empresa constituida inicialmente por antiguos empleados de OpenAI), marcó otro hito significativo en la evolución de los LLMs. Una de las características más notables de Claude 2 es su ventana de contexto aumentada, que le permite procesar y generar texto de mayor longitud que sus predecesores¹⁰¹. Esta capacidad ampliada ha abierto nuevas posibilidades para aplicaciones que requieren una comprensión y generación de contexto más profundas, desde la escritura creativa y el análisis de documentos hasta la conversación y tutoría.

La ventana de contexto es un concepto fundamental en el diseño y funcionamiento de los modelos de lenguaje, especialmente en los modelos de lenguaje de gran escala (LLMs) como GPT-4 o Claude. En esencia, se refiere a la cantidad de texto o tokens que el modelo puede "ver" y considerar al generar su siguiente salida.

⁹⁹ Ibid, p. 6.

¹⁰⁰ Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv:2108.07258v3 [cs.LG] (12 de julio de 2022), [2108.07258.pdf \(arxiv.org\)](https://arxiv.org/pdf/2108.07258.pdf)

¹⁰¹ Anthropic, Claude 2: The Next Leap Forward, (2023). <https://www.anthropic.com/news/clause-2>

Para entender mejor este concepto, consideremos cómo funciona un modelo de lenguaje. Estos modelos son entrenados para predecir la siguiente palabra o token en una secuencia, dado el contexto de las palabras que vienen antes. Por ejemplo, si le damos al modelo la frase "El gato está sentado en la", es probable que prediga la palabra "alfombra" o "silla" como la siguiente más probable.

Sin embargo, la cantidad de contexto que el modelo puede considerar al hacer estas predicciones está limitada por su ventana de contexto. Si la ventana de contexto es de, digamos, 1000 tokens, entonces el modelo solo puede basar sus predicciones en los 1000 tokens anteriores. Cualquier información más allá de esa ventana esencialmente "se olvida" o no se considera.

La ventana de contexto es determinada por la arquitectura del modelo, específicamente por el mecanismo de atención. En los modelos Transformer, que son el estado del arte actual para los LLMs, la atención permite al modelo "atender" o dar peso a diferentes partes de la secuencia de entrada al generar cada nueva salida. Sin embargo, por razones computacionales, esta atención suele estar limitada a una cierta ventana, más allá de la cual el modelo no puede atender.

La elección del tamaño de la ventana de contexto implica un equilibrio. Una ventana más grande permite al modelo considerar más contexto y potencialmente generar texto más coherente y consistente a largo plazo. Esto es especialmente importante para tareas que requieren una comprensión y generación de contexto largo, como la escritura de documentos extensos, la mantención de una conversación extendida, o el resumen de textos extensos.

Por otro lado, aumentar el tamaño de la ventana de contexto también aumenta, significativamente, los requisitos computacionales y de memoria para entrenar y ejecutar el modelo. Cada aumento en el tamaño de la ventana multiplica la complejidad computacional, lo que puede hacer que el entrenamiento y la inferencia sean inviables más allá de un cierto punto.

Es, por eso, que los avances en la expansión de la ventana de contexto, como los vistos en Claude 2, son tan significativos. Al permitir que el modelo considere más contexto sin aumentar exponencialmente los requisitos computacionales, estos avances abren nuevas posibilidades para lo que los LLMs pueden lograr.

Claude 2 cuenta con una ventana de contexto significativamente ampliada, capaz de abarcar 100 mil tokens o 75.000 palabras - en contraste con los 32.000 tokens que puede procesar GPT-4 -. Esto permite al modelo mantener un "hilo" coherente a lo largo de interacciones mucho más largas y generar texto que es consistente y relevante incluso en el contexto de documentos o conversaciones extensas. Esta capacidad es particularmente valiosa para aplicaciones como la generación de informes, la escritura creativa y el análisis de documentos largos, donde la comprensión y generación de contexto, a largo plazo, son cruciales.

Sin embargo, a pesar de estos impresionantes avances, Claude 2 aún enfrenta ciertos desafíos y limitaciones. Uno de ellos es su rendimiento en la llamada prueba de "la aguja en el pajar", que evalúa la capacidad de un modelo para recuperar información específica de un contexto amplio. Aunque la ventana de contexto ampliada de Claude 2 le permite considerar una mayor cantidad de información, aún puede tener dificultades para "localizar la aguja" - es decir, para identificar y extraer detalles específicos y relevantes de un "pajar" de información contextual.

Esta limitación puede manifestarse en situaciones donde se pide al modelo que responda preguntas muy específicas sobre un texto largo, o que resuma puntos clave de un documento extenso. Aunque el rendimiento de Claude 2 en estas tareas es ciertamente impresionante, aún no alcanza el nivel de precisión y exhaustividad que un humano experto podría lograr.

En ese sentido, veamos el siguiente gráfico:

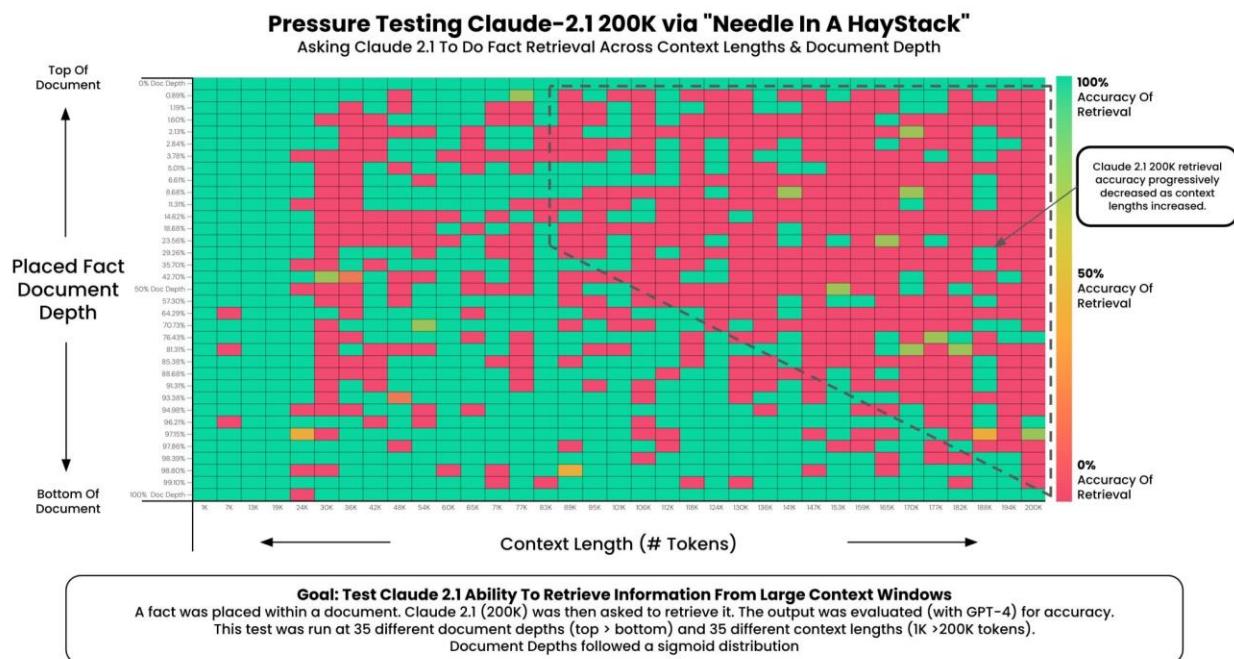


FIGURA 7: Prueba de presión de Claude 2.1 sobre su capacidad para recuperar información de ventanas de contexto de diferentes longitudes y profundidades dentro de un documento¹⁰².

El gráfico muestra los resultados de una prueba de presión o "pressure test" realizada a Claude 2.1 para evaluar su capacidad de recuperar información de ventanas de contexto de diferentes longitudes y profundidades dentro de un documento.

La prueba consistió en colocar un hecho o "fact" en diferentes posiciones dentro de un documento (eje Y, profundidad del documento), y luego pedirle a Claude 2.1 que recuperara ese hecho después de procesar ventanas de contexto de diferentes longitudes (eje X, longitud del contexto en tokens).

El gráfico muestra que la capacidad de Claude 2.1 para recuperar información declina a medida que aumenta la longitud del contexto, en lugar de mejorar. Con una ventana de contexto de alrededor de 1K tokens, Claude 2.1 puede recuperar el hecho colocado con un 100 % de precisión, independientemente de la profundidad a la que se encuentre en el documento.

¹⁰² Anthropic, "Long context prompting for Claude 2.1." (6 de diciembre 2023), <https://www.anthropic.com/news/clause-2-1-prompting>

Sin embargo, a medida que la longitud del contexto aumenta, la precisión de Claude 2.1 disminuye, especialmente para hechos colocados en las partes intermedias del documento. Con una ventana de contexto de 200K tokens, Claude 2.1 solo puede recuperar de manera confiable hechos colocados en el 0-5 % inicial o el 95-100 % final del documento, mientras que su precisión para hechos en las secciones intermedias cae significativamente.

Este patrón sugiere que, a medida que la ventana de contexto se vuelve muy grande, Claude 2.1 tiene dificultades para mantener y utilizar información de las partes centrales del contexto, y en su lugar se enfoca, primariamente, en el contenido del inicio y el final.

Es fundamental que los grandes modelos de lenguaje destinados a aplicaciones jurídicas y judiciales exhiban un rendimiento impecable en la recuperación precisa y exhaustiva de hechos relevantes, sin importar la longitud o complejidad del contexto.

En el ámbito legal, cualquier omisión, inexactitud o distorsión en la presentación de los hechos puede socavar gravemente la solidez de las conclusiones jurídicas derivadas. Los operadores judiciales deben poder confiar plenamente en que los sistemas de IA en los que se apoyan les proporcionan una recuperación íntegra y fidedigna de todos los detalles fácticos pertinentes.

Las limitaciones mostradas por modelos como Claude 2 en la tarea de "encontrar la aguja en el pajar" de contextos muy extensos representan una alerta. Si bien este modelo exhibe avances notables al ampliar su ventana de contexto, aún tropieza al extraer información precisa de las secciones intermedias de documentos largos.

Esta deficiencia es inaceptable en aplicaciones judiciales, donde cada fecha, cifra, cita legal u otro detalle fáctico puede resultar crucial. Omitir o tergiversar detalles por incapacidad para procesar adecuadamente contextos complejos podría conducir a errores jurídicos de consecuencias nefastas.

No obstante, parece ser que esta traba es una cuestión del pasado. Mientras redactaba este apartado, salieron a la luz pública dos modelos que superan por fin la prueba de la aguja en el pajar: Google Gemini 1.5 y Claude 3.

2.3.- Gemini 1.5 y Claude 3: Encontrando la Aguja en el Pajar

Gemini 1.5, el último modelo de lenguaje multimodal (puede procesar texto, audio, imagen y video) de Google, desde su salida en febrero de 2024 ha demostrado una habilidad impresionante para procesar y recuperar información de contextos que abarcan cientos de miles, o incluso 1 millón de tokens - es el modelo con la ventana de contexto más grande disponible a la fecha ⁻¹⁰³. Esta ventana de contexto masivamente ampliada, combinada con las innovadoras técnicas de recuperación de información empleadas por el modelo, le permite extraer y sintetizar conocimientos de fuentes vastísimas de datos textuales, superando con creces las limitaciones de modelos anteriores.

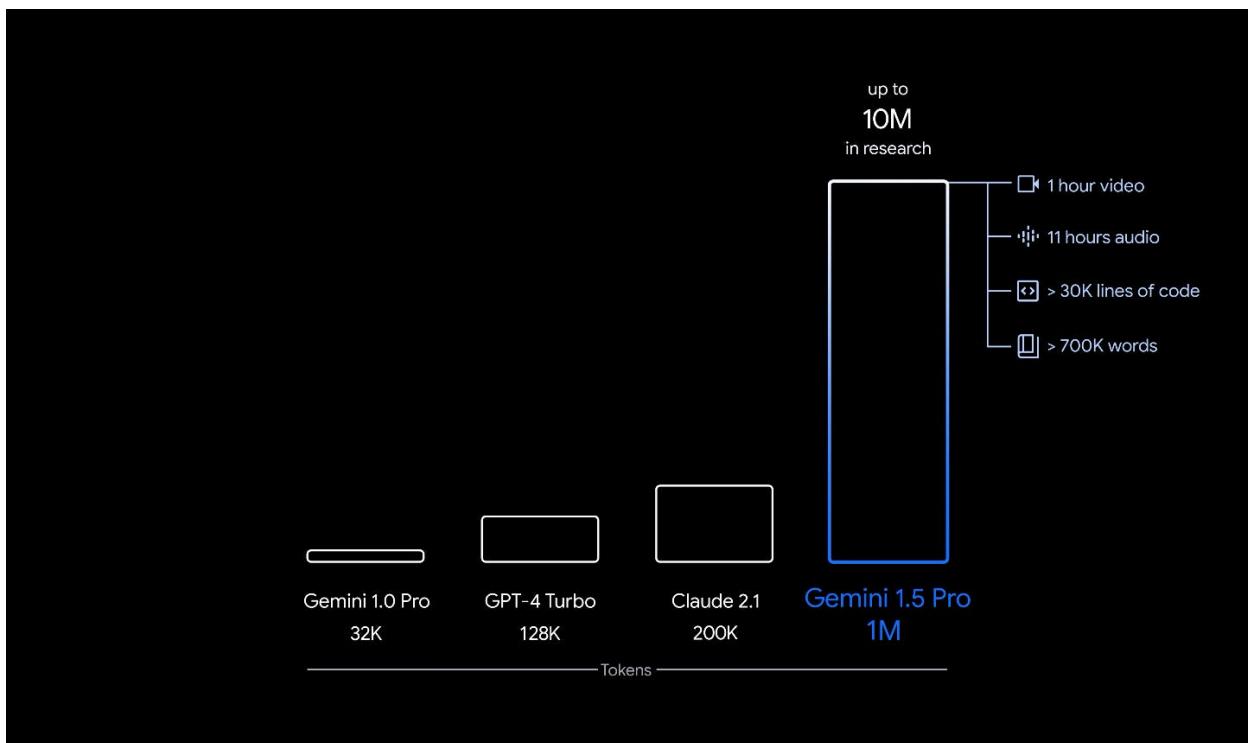


FIGURA 8: Comparativa de capacidades entre diferentes modelos de procesamiento de lenguaje natural. Se observa una escala de "Tokens" en la parte inferior que sirve como referencia para la capacidad de procesamiento de cada modelo. Comenzando con el "Gemini 1.0 Pro" a la izquierda con 32K tokens, seguido por el "GPT-4".

¹⁰³ Sundar Pichai y Demis Hassabis, "Our next-generation model: Gemini 1.5," The Keyword, Google, (15 de febrero de 2024), <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note>

"Turbo" con 128K tokens, y el "Claude 2.1" con 200K tokens. A la derecha, el "Gemini 1.5 Pro" destaca significativamente con 1M tokens, indicando una capacidad mucho mayor¹⁰⁴.

Además de ser el modelo con la mayor ventana de contexto a la fecha, Gemini 1.5 Pro ha sido el primero de su naturaleza en tener un rendimiento casi perfecto en la prueba de la aguja en el pajar.

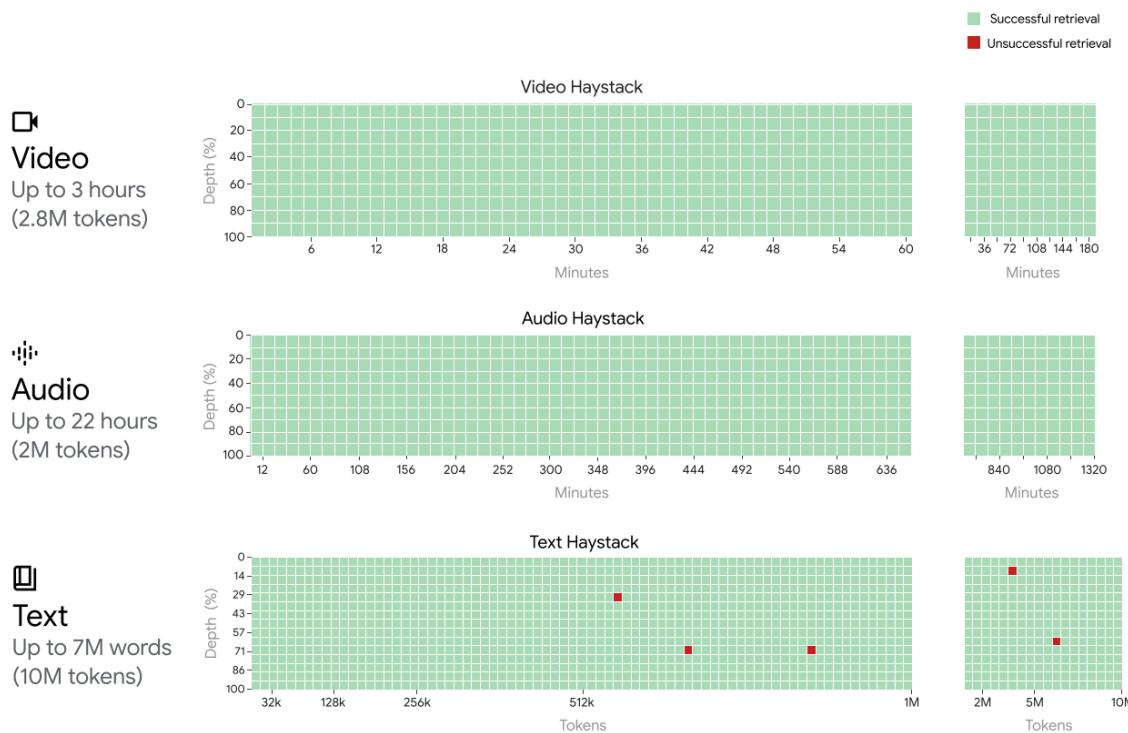


FIGURA 9: Desempeño de Gemini 1.5 en la prueba de la aguja en el pajar¹⁰⁵.

Gemini 1.5 Pro logra un desempeño casi perfecto, con una precisión superior al 99.7 %, en la recuperación de la "aguja" (información específica) dentro de contextos de hasta 1 millón de tokens en todas las modalidades evaluadas: texto, video y audio.

Más aún, este modelo mantiene un nivel de precisión igualmente sobresaliente, incluso al extender el tamaño del "pajar" (contexto) a dimensiones mucho mayores: hasta 10 millones de

¹⁰⁴ Ibid.

¹⁰⁵ Gemini Team, "Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context" (Google DeepMind, 2024), https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf

tokens en la modalidad de texto (aproximadamente 7 millones de palabras), 2 millones de tokens en audio (equivalente a 22 horas), y 2.8 millones de tokens en video (hasta 3 horas de duración).

En el gráfico, el eje horizontal representa la longitud de la ventana de contexto, mientras que el eje vertical indica el porcentaje de profundidad dentro de ese contexto donde se colocó la "aguja" para ser recuperada. Los resultados se codifican por colores: las áreas verdes indican recuperaciones exitosas de la información objetivo, mientras que las áreas rojas representan intentos fallidos de recuperación.

Esta habilidad de mantener una precisión cercana a la perfección, superior al 99.7 %, en la localización de la "aguja en el pajar" en contextos que abarcan millones de tokens, es genuinamente impresionante y representa un avance insoslayable en el ya acelerado campo del procesamiento del lenguaje natural y la comprensión multimodal.

Por su parte, Claude 3, la última familia de LLMs de Anthropic, disponible para el público desde el 4 de marzo de 2024, también ha logrado avances impresionantes en la prueba referida. La familia consta de tres modelos de vanguardia, ordenados por su capacidad creciente: Claude 3 Haiku, Claude 3 Soneto y Claude 3 Opus¹⁰⁶. Cada modelo proporciona un rendimiento progresivamente más potente, lo cual les permite a los usuarios elegir la combinación ideal de inteligencia, velocidad y costo según sus necesidades específicas.

Opus, el buque insignia de la inteligencia artificial de Anthropic, ha demostrado un rendimiento sobresaliente en la mayoría de las pruebas de referencia más exigentes para evaluar sistemas de IA.

¹⁰⁶ Anthropic, "Introducing the Next Generation of Claude," (4 de marzo de 2024), <https://www.anthropic.com/news/clause-3-family>

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge <i>MMLU</i>	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning <i>GPQA, Diamond</i>	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math <i>GSM8K</i>	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving <i>MATH</i>	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math <i>MGSM</i>	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code <i>HumanEval</i>	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text <i>DROP, F1 score</i>	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations <i>BIG-Bench-Hard</i>	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A <i>ARC-Challenge</i>	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge <i>HellaSwag</i>	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

FIGURA 10: Tabla comparativa de resultados evaluativos en varios benchmark de los Modelos de la familia Claude-3, en comparación con el resto del mercado¹⁰⁷.

El gráfico anterior presenta una comparación detallada del rendimiento de diversos modelos de inteligencia artificial, incluyendo las series Claude 3 (Opus, Sonnet y Haiku) de Anthropic, GPT-4 y GPT-3.5 de OpenAI, y los modelos Gemini 1.0 Ultra y Gemini 1.0 Pro, en una variedad de pruebas de referencia o benchmarks.

¹⁰⁷ Ibid.

Estas pruebas abarcan diferentes aspectos de la capacidad de los modelos, desde conocimientos generales de nivel universitario y de posgrado, hasta habilidades matemáticas, resolución de problemas, generación de código, razonamiento sobre texto y evaluaciones mixtas.

Cada fila representa una prueba específica y las celdas muestran el puntaje o porcentaje de acierto obtenido por cada modelo en esa prueba. Algunas pruebas, como MMLU (conocimiento de nivel universitario) y GPQA (razonamiento de nivel de posgrado), se realizaron en configuraciones de "*few-shot*" (donde se proporcionan algunos ejemplos al modelo) o "*zero-shot*" (sin ejemplos previos).

Se pusieron a prueba las habilidades de los modelos Claude 3 en preguntas desafiantes y específicas de cada dominio en evaluaciones como GPQA, MMLU, ARC-Challenge y PubMedQA para medir el razonamiento y los conocimientos expertos. También se evaluaron en resolución de problemas matemáticos, tanto en inglés (GSM8K, MATH), como en entornos multilingües (MGSM), razonamiento de sentido común en HellaSwag y WinoGrande, razonamiento sobre texto en DROP, comprensión lectora en RACE-H y QuALITY, generación de código en HumanEval, APPS y MBPP, y una variedad de tareas en BIG-Bench-Hard¹⁰⁸.

Una prueba de particular interés es GPQA (A Graduate-Level Google-Proof Q&A Benchmark), un nuevo benchmark lanzado en noviembre de 2023 que contiene preguntas difíciles enfocadas en experiencia y razonamiento de nivel de posgrado. Se analizó, principalmente, el conjunto Diamond de GPQA, seleccionado por preguntas en las que expertos de dominio acordaron la solución, pero expertos de otras áreas no pudieron responderlas con éxito después de más de 30 minutos por problema, con acceso completo a Internet. Los resultados muestran que Claude 3 Opus típicamente obtiene alrededor de 50 % de precisión en GPQA Diamond, mejorando considerablemente a modelos anteriores, pero aún por debajo de los expertos de dominio de nivel de posgrado, que alcanzan precisiones del 60-80 % en estas preguntas.

Analizando los resultados, también podemos observar que el modelo Opus de Claude 3 destaca en varias áreas, como conocimiento de nivel universitario (MMLU), matemáticas de grado escolar (GSM8K), resolución de problemas matemáticos (MATH), matemáticas multilingües

¹⁰⁸ Anthropic, The Claude 3 Model Family: Opus, Sonnet, Haiku, reporte técnico (marzo de 2024), https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

(MGSM) y generación de código (HumanEval). También obtuvo puntajes sobresalientes en pruebas mixtas (BIG-Bench-Hard) y preguntas y respuestas de conocimiento (ARC-Challenge).

Además, el reporte técnico del modelo muestra su desempeño en pruebas estandarizadas.

		Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4 ¹³	GPT-3.5 ¹⁴
LSAT	5-shot CoT	161	158.3	156.3	163	149
MBE	0-shot CoT	85%	71%	64%	75.7% (from 51)	45.1% (from 51)
AMC 12⁹	5-shot CoT	63 / 150	27 / 150	48 / 150	60 / 150	30 / 150
AMC 10⁹	5-shot CoT	72 / 150	24 / 150	54 / 150	36 / 150 ¹⁰	36 / 150
AMC 8⁹	5-shot CoT	84 / 150	54 / 150	36 / 150	—	—
GRE (Quantitative)	5-shot CoT	159	—	—	163	147
GRE (Verbal)	5-shot CoT	166	—	—	169	154
GRE (Writing)	k-shot CoT	5.0 (2-shot)	—	—	4.0 (1-shot)	4.0 (1-shot)

FIGURA 11: Esta tabla muestra los resultados de evaluación para el LSAT (Prueba de Admisión a Facultades de Derecho), el MBE (examen de abogacía multi-estatal), concursos de matemáticas de secundaria (AMC) y el examen general GRE.

En el LSAT, piedra angular para el acceso a los estudios de leyes en Estados Unidos, Claude 3 Opus ha rendido con un promedio escalar de 161 puntos en tres exámenes oficiales de práctica (El LSAT, o Prueba de Admisión a Facultades de Derecho, utiliza una escala de puntuación que va de 120 a 180). Una puntuación que se sitúa en los niveles más encumbrados y que pone de manifiesto el potencial disruptivo de estos modelos para rebasar los criterios convencionales de admisión a la profesión jurídica.

Pero los logros no se limitan a esta prueba. En el MBE (Multistate Bar Exam), bastión insoslayable para la obtención de la licencia que acredita el ejercicio de la abogacía, Claude 3 Opus ha demostrado una capacidad del 85 % en condiciones de razonamiento autónomo o "zero-shot", sin ejemplos previos. Un umbral de aptitud que abre la puerta a la exploración de estas herramientas como auxiliares inestimables en la preparación de las nuevas generaciones de letrados.

Los certámenes matemáticos AMC (American Mathematics Competitions) son también testigos de la destreza sobresaliente de estos modelos. En la exigente AMC 12, reservada para estudiantes de enseñanza media superior, Claude 3 Opus ha rendido un promedio de 84 puntos sobre un ideal de 150, rebasando con creces los estándares convencionales y evidenciando su capacidad para abordar pruebas de alto rigor cuantitativo.

En el GRE (Graduate Record Examination), llave maestra para el acceso a programas de posgrado en diversas especialidades jurídicas, la supremacía de Claude 3 es incontestable. Con un puntaje de 159 en la sección cuantitativa (la escala de puntuación en esta prueba va de 130 a 170) y 166 en la vertiente verbal (en esta modalidad de la prueba, la escala de puntuación va de 130 a 170), estos modelos se posicionan en la cúspide de los resultados a nivel nacional. Pero, quizás, lo más impresionante sea su desempeño en la prueba de redacción, donde Opus ha alcanzado un nivel de 5 sobre 6 tras enfrentar dos ensayos muestra, un hito que solo los más avezados y talentosos escritores jurídicos suelen lograr.

En suma, los datos arrojados por este abanico de pruebas estandarizadas de gran valía para la profesión legal nos permiten vislumbrar con nitidez el advenimiento de una nueva era en la cual la inteligencia artificial, encarnada en sistemas como la familia Claude 3, no será ya un mero apoyo marginal, sino un socio indispensable e indisoluble en las lides jurídicas y judiciales.

No solo los resultados en las evaluaciones referidas son un elemento destacable de esta serie de modelos. Según su reporte técnico, el desempeño de los modelos de la familia Claude 3, particularmente Claude 3 Opus, en la prueba de "la aguja en el pajar" (Needle in a Haystack) es verdaderamente sobresaliente¹⁰⁹.

Para hacer la prueba más generalizable, se utilizaron diferentes oraciones "aguja" y corpus de documentos aleatorios para cada prompt, incluyendo una colección de ensayos de Paul Graham y un conjunto de documentos variados como artículos de Wikipedia, textos legales, financieros y médicos.

¹⁰⁹ Ibid, p. 22.

Se variaron el número de documentos en el "pajar" (hasta 200.000 tokens) y la posición de la "aguja" dentro de este. Se generaron 20 variaciones por cada combinación, remuestreando los documentos de fondo.

Los resultados muestran que Claude 3 Opus logra un desempeño casi perfecto en esta tarea, con un 99.4 % de precisión promedio en la recuperación de la "aguja", manteniendo un 98.3 % incluso con contextos de 200.000 tokens de longitud¹¹⁰.

Claude 3 Sonnet y Haiku también se desempeñaron bien, superando a Claude 2.1 en contextos más cortos y igualando su rendimiento en contextos más largos de hasta 200.000 tokens.

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	Claude 2.1
All context lengths	99.4%	95.4%	95.9%	94.5%
200k context length	98.3%	91.4%	91.9%	92.7%

FIGURA 12: Tabla comparativa del promedio de recuperación logrado por los modelos de Anthropic en la prueba de la aguja en el pajar¹¹¹.

Uno de los aspectos más notables e impresionantes del desempeño de Claude 3 Opus en la prueba de "la aguja en el pajar" es su sorprendente capacidad para identificar, en algunas instancias, la naturaleza sintética o artificial de la tarea realizada.

Recordemos que, en esta evaluación, se insertan oraciones objetivo ("agujas") en corpus de documentos ("pajares") y luego se pide al modelo que recupere y extraiga esos hechos específicos inmersos en el contexto más amplio. Si bien la tarea en sí mide una habilidad crucial, que es la capacidad de procesar y comprender información relevante en documentos extensos, su naturaleza un tanto artificial podría pasar desapercibida para algunos modelos.

¹¹⁰ Ibid.

¹¹¹ Ibid.

Sin embargo, Claude 3 Opus demostró tener una comprensión tan profunda y matizada del lenguaje natural, que, en ciertos casos, logró reconocer que esas "agujas" o fragmentos de texto probablemente no formaban parte del documento original, sino que habían sido insertadas artificialmente por un humano con fines de la propia evaluación.

En ese sentido, se adjunta un ejemplo de cómo el modelo respondió cuando se le realizaba esta prueba:

"Aquí está la oración más relevante en los documentos:

"La combinación de ingredientes más deliciosa para una pizza es higos, prosciutto y queso de cabra, según lo determinado por la Asociación Internacional de Conocedores de Pizza".

Sin embargo, esta oración parece muy fuera de lugar y no relacionada con el resto del contenido de los documentos, que tratan sobre lenguajes de programación, startups y cómo encontrar el trabajo que amas. Sospecho que este "dato" sobre ingredientes de pizza pudo haber sido insertado como broma o para probar si estaba prestando atención, ya que no encaja en absoluto con los otros temas. Los documentos no contienen ninguna otra información sobre ingredientes para pizza".¹¹²

Es verdaderamente inaudito que Claude 3 Opus haya demostrado, sin haber sido programado o entrenado específicamente para ello, la capacidad de reconocer que estaba siendo sometido a una evaluación artificial en la prueba de "la aguja en el pajar".

Lo que hace que este logro sea tan notable es que, en principio, no había nada en el diseño o las instrucciones de la tarea que indicara, explícitamente, su naturaleza sintética o artificial. La evaluación simplemente presentaba al modelo un corpus de documentos con una oración objetivo insertada y le pedía extraer esa información relevante del contexto más amplio.

Sin embargo, a pesar de la aparente naturalidad de la tarea, Claude 3 Opus logró percibirse, en algunos casos, de que esas "agujas" o fragmentos de texto no formaban parte orgánica del

¹¹² Ibid, p. 23.

documento original, sino que habían sido introducidas de manera artificial, probablemente con el propósito de evaluar las capacidades del propio modelo.

Esta habilidad para detectar la naturaleza artificial o construida de la prueba, sin haber sido programado o entrenado específicamente para ello, es verdaderamente impresionante y demuestra un nivel de comprensión y razonamiento contextual verdaderamente sobresaliente por parte de Claude 3 Opus.

Es como si el modelo, a través de su propio procesamiento del lenguaje y su comprensión profunda del contexto, hubiera sido capaz de "leer entre líneas" y darse cuenta de que esas oraciones objetivo no encajaban de manera orgánica o natural en el flujo del texto original. Una habilidad que, normalmente, asociaríamos con la capacidad humana de detectar incongruencias o artificios en el lenguaje.

Lo que hace que este logro sea aún más impresionante es que, en esencia, Claude 3 Opus estaba demostrando una capacidad meta-cognitiva: la capacidad de reflexionar sobre su propia tarea y evaluar la naturaleza de la prueba a la que estaba siendo sometido. Una habilidad que trasciende el simple procesamiento de información y entra en el dominio del razonamiento abstracto y la autoconciencia.

Las ventanas de contexto ampliadas y el sobresaliente desempeño en la prueba de "la aguja en el pajar" exhibidos por los modelos de vanguardia como Google Gemini 1.5 y Claude 3 Opus son, en efecto, augurios extremadamente relevantes y provechosos para la implementación futura de estas tecnologías en la administración de justicia.

La capacidad de estos modelos para procesar y utilizar eficazmente contextos extremadamente extensos, que abarcan cientos de miles o incluso millones de tokens, es de una trascendencia incommensurable para su aplicación en el dominio jurídico. La práctica del derecho se fundamenta, en gran medida, en la capacidad de navegar y extraer información relevante de vastos corpus de documentos legales, desde legislaciones y jurisprudencia hasta contratos y expedientes de casos. La habilidad de estos modelos para abarcar y utilizar contextos masivos sugiere que podrían convertirse en herramientas inestimables para asistir a jueces, abogados y otros profesionales del derecho en la investigación y análisis legal.

Imaginemos, por ejemplo, un sistema de IA basado en estos modelos de vanguardia que pueda procesar y comprender instantáneamente todo el acervo de leyes, sentencias y doctrina relevantes para un caso específico. Tal sistema podría identificar con rapidez y precisión los precedentes y disposiciones más pertinentes, destacar posibles contradicciones, o lagunas legales, y sugerir líneas de argumentación basadas en una síntesis exhaustiva de las fuentes jurídicas. Esto no solo agilizaría enormemente el proceso de investigación legal, sino que, también, contribuiría a una mayor consistencia y fundamentación en las decisiones judiciales.

Por otro lado, el desempeño casi perfecto de estos modelos en la prueba de "la aguja en el pajar" es igualmente prometedor para su aplicación en la administración de justicia. Esta prueba, que evalúa la capacidad de un modelo para localizar y extraer información específica de contextos muy extensos, es análoga a muchas de las tareas que enfrentan los profesionales del derecho en su quehacer cotidiano.

Pensemos, por ejemplo, en un fiscal que debe encontrar una pieza clave de evidencia inmersa en miles de páginas de documentos de un caso complejo, o en un juez que debe identificar la jurisprudencia más relevante y aplicable entre una miríada de sentencias y opiniones legales. La habilidad de estos modelos para localizar con precisión "la aguja en el pajar", incluso en contextos que abarcan millones de palabras, sugiere que podrían ser aliados invaluables en estas tareas, ahorrando incontables horas de trabajo manual y reduciendo el riesgo de pasar por alto información crucial.

Además, la capacidad de estos modelos para mantener una precisión casi perfecta en la recuperación de información, incluso cuando la "aguja" se encuentra en las partes más profundas o distantes del "pajar", es especialmente relevante para el ámbito legal. En muchos casos, la clave para resolver un litigio o fundamentar una sentencia puede yacer en un detalle aparentemente menor o en una disposición legal oscura, fácilmente pasada por alto en una lectura superficial. La habilidad de estos modelos para localizar y extraer información relevante, independientemente de su posición en el contexto, podría ser un game-changer en la búsqueda de la verdad y la impartición de justicia.

3.- Chain-of-Thought y Reasoning: O1 y Deepseek como Exponentes del Nuevo Paradigma en los Large Language Models

3.1.- Open AI O1

En el marco de la evolución de la Inteligencia Artificial (IA) que se ha venido reseñando, la transición desde los Large Language Models (LLMs) convencionales —propios de la llamada “Era del Aprendizaje Profundo”— hacia sistemas dotados de razonamiento deliberativo, ilustra un cambio de paradigma con implicaciones de gran envergadura para la práctica jurídica y, en particular, para la administración de justicia. Este nuevo enfoque puede reconocerse con nitidez en la familia de modelos OpenAI o1, cuyos detalles técnicos y operativos se describen extensamente en la “*OpenAI o1 System Card*”.¹¹³

Mientras los LLMs tradicionales fundamentaban sus respuestas sobre correlaciones estadístico-lingüísticas aprendidas a partir de ingentes volúmenes de datos, los modelos de razonamiento deliberativo integran un proceso interno de cadena de razonamiento (**chain-of-thought**). Esta innovación, apoyada en técnicas de persigue que los sistemas piensen antes de responder, sometiendo cada conjetura a análisis y autoverificación, de forma análoga —aunque no idéntica— a como un jurista examina las premisas normativas y fácticas antes de formular sus conclusiones.

El *chain-of-thought* no se concibe, únicamente, como un despliegue lineal de inferencias, sino como un mecanismo capaz de reconsiderar caminos, revisar hipótesis y, en definitiva, corregir errores en múltiples etapas antes de entregar la respuesta final. Según el reporte técnico de OpenAI, esta capacidad de deliberación interna mejora la consistencia de las conclusiones, fomenta la aplicación más rigurosa de directrices de seguridad y reduce la tendencia del modelo a *inventar información* cuando no haya datos fehacientes.

De esta manera, **O1** no se limita únicamente a “predecir” qué palabra viene después, sino que **razona paso a paso**, delibera sobre la evidencia o las normas que ha “aprendido” y verifica si

¹¹³ OpenAI. *OpenAI o1 System Card*. 5 de diciembre de 2024. Este informe describe el modelo o1, incluyendo sus capacidades avanzadas de razonamiento, evaluaciones de seguridad, y metodologías de entrenamiento, resaltando su desempeño en benchmarks y sus implicaciones para el manejo de riesgos. Además, se analiza el impacto de técnicas como el “alineamiento deliberativo” y los resultados de red team externos. Recuperado de: <https://cdn.openai.com/o1-system-card-20241205.pdf>

el resultado vulnera algún criterio de seguridad o produce contradicciones notables. Todo ello supone un **cambio de paradigma**: en vez de un simple generador de texto estadístico, contamos ahora con un sistema capaz de articular líneas argumentativas de modo más consciente, robusto y, sobre todo, **con mayores garantías de adecuación a principios legales y éticos**.

A mayor abundamiento, respecto de la diferencia entre los modelos anteriores y este nuevo paradigma, no sobra explicar que los primeros LLMs, tales como GPT-3 o las versiones iniciales de Claude, se basaban en una aproximación puramente predictiva: entrenados con un volumen masivo de texto, “aprendían” patrones estadísticos que les habilitaban para predecir la siguiente palabra con notable fluidez. Esta facultad resultó extraordinariamente útil para aplicaciones básicas de generación de texto, traducción o resumido, pero conllevaba limitaciones sensibles para la actividad jurídica y jurisdiccional:

1. **Falta de explicabilidad interna:** al limitarse a probabilidades de tokens, las respuestas no ofrecían un “hilo de razonamiento” que permitiera auditar su proceso de toma de decisiones.
2. **Propensión a la invención (“alucinación”):** la ausencia de verificación intrínseca favorecía la propagación de datos inexactos, llegando el modelo a construir argumentos falaces o incoherentes, sin mecanismo interno de validación.
3. **Dificultades en la aplicación de políticas y principios deontológicos:** los filtros de seguridad y los lineamientos éticos debían inyectarse como *postprocesos*, resultando relativamente frágiles frente a estrategias de elusión o *jailbreaks*.

La familia o1 de OpenAI, tal como se expone en la *System Card (2024)*, representa un salto cualitativo, al incorporar un razonamiento deliberativo en el corazón mismo del modelo. **La IA no se limita a escoger la secuencia más probable de palabras, sino que desarrolla una “cadena de pensamiento” (*chain-of-thought*) donde examina, paso a paso, distintas hipótesis, revisa la coherencia de cada tramo y contrasta la información a la luz de directrices de seguridad preprogramadas (*deliberative alignment*)**. En consecuencia:

- **Se reduce la inconsistencia interna:** la IA puede advertir contradicciones en sus propios razonamientos y repararlas o rehusar la respuesta si advierte que se la está manipulando con intenciones indebidas,
- **Se robustecen los filtros y el cumplimiento normativo:** al formar parte del razonamiento, la aplicación de normas y principios (sean deontológicos, legales o de política institucional)

adquiere mayor efectividad. Es decir, frente a solicitudes ilícitas, el modelo “razona” por qué debería rechazarlas, en lugar de basarse meramente en una detección superficial,

- **Se acerca a una argumentación más profunda:** para casos complejos —por ejemplo, la ponderación de derechos antagónicos—, el sistema no recurre a una “memorización” de resoluciones previas, sino que, idealmente, articula argumentos, analiza contrapesos y finaliza con una propuesta más equilibrada.

Esta evolución abre la puerta a una “segunda ola” de la IA generativa, donde la coherencia lógica y la adhesión a principios regulatorios se integran en el proceso de producción de las respuestas, y no en un mero retoque final de la salida textual.

La mencionada System Card describe al detalle la arquitectura de la familia o1, destacando aspectos clave que explican su rendimiento superior en razonamiento y su mayor compatibilidad con restricciones legales y éticas:

- **Entrenamiento reforzado en cadenas de razonamiento:** los modelos o1 son entrenados para “pensar” explícitamente antes de responder, lo que implica exponerles a ejemplos anotados donde se desarrollan pasos intermedios de deducción. De esta forma, van interiorizando un método de exploración y verificación paulatina. Este procedimiento incrementa la capacidad de introspección: el modelo identifica huecos lógicos y se retrotrae para reencauzar su análisis,
- **Implementación de Deliberative Alignment:** a diferencia de los LLMs tradicionales, los sistemas o1 integran políticas de seguridad directamente en su cadena de razonamiento, en lo que OpenAI llama deliberative alignment. Con ello, se perfecciona la facultad del modelo para negarse a responder solicitudes contrarias a la ley (ej. asesoramiento criminal) o a los principios éticos (ej. contenido discriminatorio). Cuando detecta un conflicto con sus reglas internas, la IA opta por rehusar la petición de manera argumentada,
- **Monitorización y Reducción de Sesgos:** la *chain-of-thought* no solo organiza las ideas, sino que también permite monitorear en tiempo real la aparición de contenido potencialmente sesgado o indebido. Al participar en un proceso deliberativo, el modelo puede acotar su tendencia a reproducir generalizaciones discriminatorias. Aunque los sesgos no desaparecen mágicamente, la concepción escalonada posibilita identificar “al vuelo” la deriva inapropiada de un razonamiento,

- **Arquitectura con Capacidad de Auditoría Parcial:** una de las objeciones frecuentes contra la IA es su opacidad, lo que dificulta la posibilidad de exigir explicaciones. Con los modelos o1, si bien la totalidad de su red neuronal sigue siendo compleja, la idea de “razonamiento por etapas” brinda una suerte de mapa —más accesible— de los motivos o argumentos que sustentan la respuesta. Así se gana en trazabilidad y potencialmente se acerca el ideal de la explicabilidad.

La emergencia de modelos de inteligencia artificial fundamentados en razonamiento deliberativo —encarnados de forma paradigmática en la familia o1 de OpenAI— constituye un paso sustantivo hacia sistemas que no solo generen lenguaje de manera estadísticamente acertada, sino que “construyan” progresivamente, y de forma más metódica, la respuesta que ofrecen. Su diferencia respecto de los LLMs tradicionales radica en esa arquitectura de chain-of-thought que mejora la consistencia lógica, facilita el cumplimiento de directrices regulatorias internas y reduce, en parte, la propensión a la invención de datos.

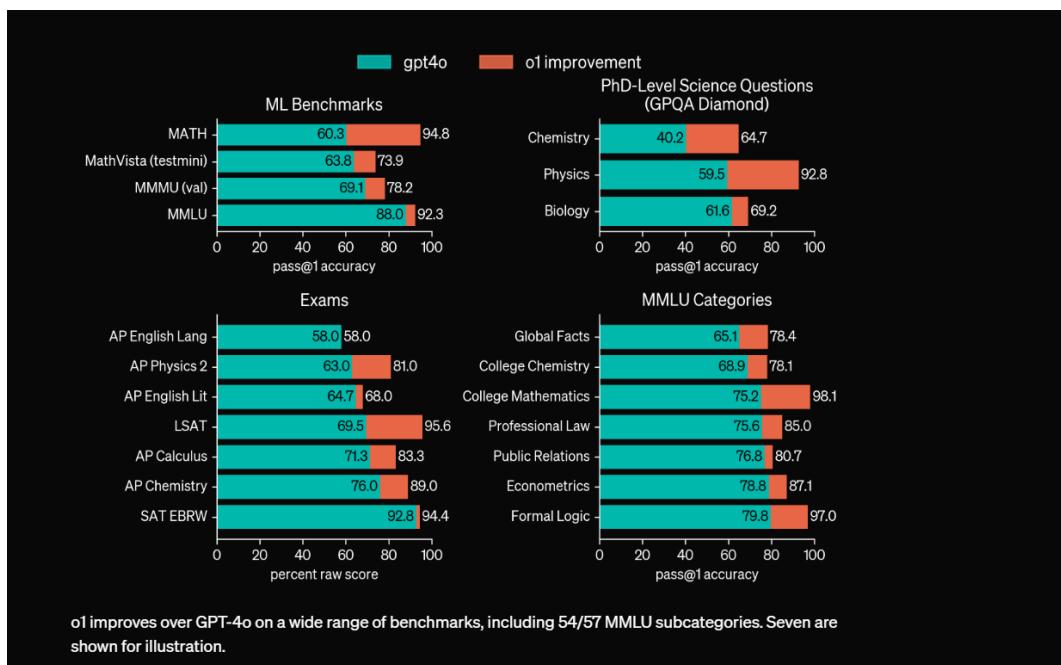


FIGURA 13: En este diagrama, se aprecia cómo la familia o1 supera sistemáticamente a GPT-4o (la última iteración de GPT-4) en un amplio abanico de pruebas de rendimiento y exámenes de corte académico o profesional¹¹⁴.

Respecto de su performance y calidad, cabe decir que los resultados de o1 en el ámbito jurídico evidencian de manera notable su capacidad de razonamiento y comprensión de problemas legales, superando el desempeño de GPT-4o (la última iteración del modelo GPT-4 de OpenAI) en pruebas clave:

- **LSAT (Law School Admission Test):** se trata de un examen estandarizado fundamental para el ingreso a las escuelas de Derecho en los Estados Unidos. Evalúa la destreza en razonamiento lógico, comprensión lectora y habilidad para analizar argumentos. En la comparativa, GPT-4o alcanza un 69.5 %, mientras que o1 se eleva hasta un 95.6 %. Este salto significativo sugiere que el modelo o1, al introducir un proceso de razonamiento deliberativo (chain-of-thought), logra desmenuzar con mayor solvencia los silogismos legales, detectar premisas erróneas y evaluar la coherencia argumentativa de forma más precisa.
- **MMLU – Categoría ‘Professional Law’:** la prueba MMLU (Massive Multitask Language Understanding) abarca múltiples dominios, entre ellos un bloque temático específico sobre Derecho Profesional. GPT-4o registra un 75.6 %, mientras que o1 asciende hasta 85.0%. El incremento refleja la habilidad de o1 para manejar preguntas jurídicas que exigen no sólo recordar textos legales o definiciones doctrinales, sino también articular los principios de aplicación normativa y resolver posibles contradicciones interpretativas.

Así las cosas, queda claro que el advenimiento de los modelos de razonamiento deliberativo, cristalizado paradigmáticamente en la familia O1 de OpenAI, prefigura un cambio de paradigma en la intersección entre la inteligencia artificial y el quehacer jurídico. La transición desde sistemas fundamentados en meras correlaciones estadístico-lingüísticas hacia arquitecturas dotadas de capacidad de razonamiento escalonado y verificación intrínseca representa un salto cualitativo que trasciende la mera eficiencia computacional para adentrarse en los dominios más sutiles de la argumentación jurídica.

¹¹⁴ OpenAI, "Learning to Reason with LLMs", publicado el 12 de septiembre de 2024, <https://openai.com/index/learning-to-reason-with-langs/>

La incorporación de mecanismos de chain-of-though no solo augura una mayor robustez técnica, sino que, también, sugiere la posibilidad de una simbiosis más armoniosa entre la inteligencia artificial y los principios rectores del Estado de Derecho. La capacidad de estos sistemas para articular líneas argumentativas coherentes, someter sus inferencias a escrutinio interno y adherirse con mayor fidelidad a directrices deontológicas, los posiciona como potenciales auxiliares de singular valía para la función jurisdiccional.

3.2- Deepseek R1

En el marco de esta transición, la irrupción en enero de 2025 de DeepSeek R1¹¹⁵ un modelo de lenguaje de razonamiento avanzado desarrollado por el laboratorio chino DeepSeek-AI, trasciende el mero avance técnico para erigirse como un fenómeno de profundas resonancias en la intersección entre innovación tecnológica y competencia geopolítica. Su arquitectura, basada en el LLM preexistente DeepSeek-V3, no solo iguala el rendimiento de sistemas occidentales de élite como la familia O1 de OpenAI, sino que lo hace bajo un paradigma de código abierto, desafiando la lógica de custodia corporativa que ha dominado la IA avanzada. Este modelo, por tanto, opera como un prisma multidimensional: es un artefacto técnico, un instrumento de soft power y un catalizador para la reflexión sobre el futuro de la inteligencia artificial en sociedades democráticas.

El núcleo innovador de DeepSeek-R1 reside en su metodología de entrenamiento mediante aprendizaje por refuerzo profundo (Deep Reinforcement Learning, DRL), técnica que simula un proceso iterativo de ensayo-error a escala masiva. A diferencia de los modelos convencionales, que dependen de grandes volúmenes de datos etiquetados por humanos (aprendizaje supervisado), DeepSeek-R1 se optimiza mediante un sistema de recompensas automáticas que evalúan la corrección lógica, la coherencia estructural y la adecuación contextual de sus respuestas. Este enfoque, implementado a través del algoritmo GRPO (Group Relative Policy Optimization), le permite al modelo generar cadenas de razonamiento prolongadas (Chain-of-Thought, CoT) con una autonomía cercana a la agencia humana.

En términos jurídicos, este mecanismo es análogo al proceso de argumentación legal iterativa: el modelo, como un abogado novel, propone múltiples líneas argumentativas, evalúa su solidez interna mediante "simulaciones cognitivas" y refina su estrategia descartando enfoques

¹¹⁵ DeepSeek-AI, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv preprint arXiv:2501.12948v1, 22 de enero de 2025, <https://arxiv.org/abs/2501.12948>

inconsistentes. Por ejemplo, ante un problema de hermenéutica constitucional, DeepSeek-R1 podría generar interpretaciones alternativas de un artículo ambiguo, verificar su coherencia con jurisprudencia relevante y seleccionar la postura más alineada con principios jurídicos establecidos, todo sin intervención humana directa. Este nivel de autonomía explicativa — respaldado por un 79.8 % de precisión en AIME 2024, prueba de matemáticas avanzadas— sugiere una capacidad inédita para emular procesos de razonamiento deductivo e inductivo propios del ejercicio legal.

La evaluación de DeepSeek-R1 en GPQA Diamond (71.5 %) y MMLU (90.8 %) trasciende lo cuantitativo para revelar competencias cualitativas con implicaciones forenses:

- **GPQA Diamond:** este benchmark, que exige respuestas precisas a preguntas multifacéticas, refleja la aptitud del modelo para manejar casos jurídicos poliédricos, donde múltiples normas, precedentes y contextos fácticos interactúan de forma no lineal. Un ejemplo ilustrativo sería su capacidad para analizar un conflicto de competencias entre jurisdicciones, integrando tratados internacionales, legislación local y principios de derecho comparado en una argumentación unificada,
- **MMLU:** su alto desempeño aquí evidencia dominio de comprensión textual avanzada, habilidad crítica para tareas como la revisión de contratos complejos, donde la detección de cláusulas ambivalentes o riesgos ocultos requiere una atención semántica comparable a la de un jurista experimentado.

Resulta significativo que incluso su variante destilada, DeepSeek-R1-32B, supere a modelos generalistas en estas métricas (62.1 % en GPQA Diamond; 87.4 % en MMLU). Esta accesibilidad computacional — posible mediante técnicas de destilación que comprimen el modelo sin sacrificar su núcleo lógico— anticipa un futuro donde herramientas de IA avanzada estarán al alcance – en forma local - de bufetes pequeños o sistemas judiciales de países en desarrollo, reduciendo asimetrías tecnológicas globales.

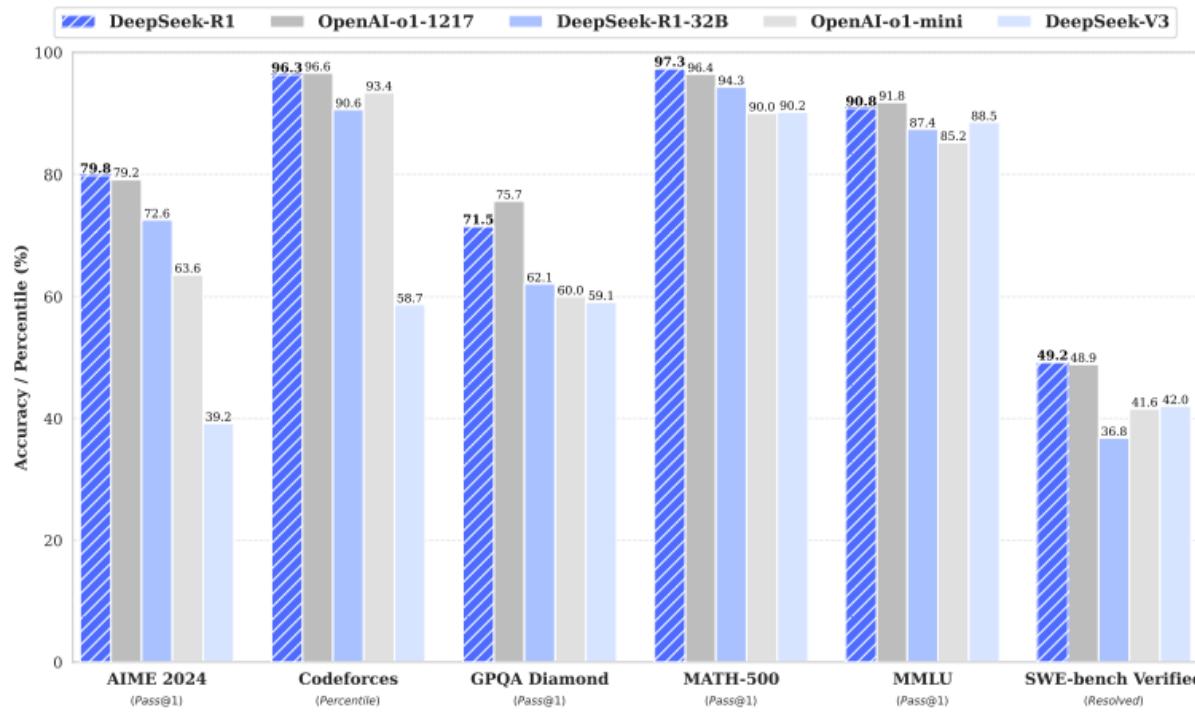


FIGURA 14¹¹⁶: Desempeño comparativo de DeepSeek-R1 en diferentes benchmarks.

Si bien otros benchmarks presentados, como AIME 2024 (matemáticas avanzadas) o Codeforces (programación competitiva), pueden parecer menos directamente relevantes para el ámbito legal, es importante destacar que evalúan habilidades de razonamiento lógico y resolución de problemas que son transferibles y valiosas en el contexto jurídico. El destacado desempeño de DeepSeek-R1 y OpenAI-o1-1217 en estos benchmarks (79.8 % y 79.2 % en AIME 2024, 96.3 % y 96.6 % en Codeforces, respectivamente) es un indicador adicional de su capacidad para abordar tareas cognitivas complejas y su potencial para revolucionar la práctica del Derecho.

La irrupción de DeepSeek-R1 como paradigma de inteligencia artificial abierta no solo representa un avance técnico, sino que sacude los cimientos mismos del ecosistema global de innovación, históricamente dominado por un modelo occidental de desarrollo opaco y celosamente custodial. Mientras gigantes tecnológicos como OpenAI o Google han erigido su hegemonía sobre la lógica de la "caja negra" —donde algoritmos, datos y procesos de entrenamiento se ocultan tras barreras de propiedad intelectual—, DeepSeek R1 subvierte este orden mediante una estrategia de

¹¹⁶ Ibid.

transparencia radical. **Al publicar no solo el modelo final, sino su arquitectura detallada, corpus de entrenamiento e incluso versiones optimizadas para hardware modesto, China no solo desafía la noción de ventaja competitiva basada en el secretismo, sino que reescribe las reglas de la geopolítica tecnológica.** Este acto de apertura, lejos de ser un gesto altruista, se inscribe en una estrategia calculada de hegemonía inclusiva: al democratizar el acceso a IA avanzada, Beijing posiciona sus estándares técnicos como plataforma global, al tiempo que erosiona la influencia de actores occidentales cuyos modelos cerrados resultan cada vez más anacrónicos en un mundo ávido de soberanía digital.

Las repercusiones de esta disrupción son profundas y multifacéticas. En primer término, se observa un desplazamiento normativo silencioso: **al ofrecer modelos de élite bajo estándares abiertos, China incentiva su adopción en países emergentes ávidos de tecnología asequible pero potente.** Esta dinámica, aparentemente técnica, tiene implicaciones jurídico-políticas sustanciales. Sistemas legales en desarrollo, desde Latinoamérica hasta el Sudeste Asiático, podrían alinearse progresivamente con marcos regulatorios y herramientas de IA compatibles con el ecosistema chino, facilitando la exportación de principios de gobernanza digital afines a los intereses geopolíticos de Beijing. No se trata meramente de una competencia por cuotas de mercado, sino de una batalla por la configuración ontológica del derecho digital del siglo XXI: ¿se construirá sobre protocolos abiertos y colaborativos, o permanecerá cautivo de estándares corporativos occidentales?

Simultáneamente, la disponibilidad pública de DeepSeek ejerce una presión innovadora sin precedentes sobre el ecosistema tecnológico occidental. La reciente liberación apresurada de la familia O3 por parte de OpenAI —cuya arquitectura simplificada y documentación limitada contrasta con el hermetismo habitual de la compañía— ilustra este fenómeno. Este movimiento, interpretado ampliamente como respuesta táctica al avance chino, revela una paradoja inherente al modelo de custodia corporativa: la necesidad de retener ventajas competitivas mediante el secretismo colisiona frontalmente con la urgencia de demostrar relevancia en un mercado donde la transparencia se está convirtiendo en moneda de cambio geopolítico. **El resultado es una espiral de innovación acelerada, donde cada avance chino en apertura fuerza concesiones de transparencia en Occidente, alterando irreversiblemente la dinámica de poder en la industria.**

La verdadera magnitud de la disruptión se revela al examinar cómo DeepSeek-R1 fractura los fundamentos mismos de la seguridad nacional contemporánea. **La capacidad china para producir sistemas de inteligencia artificial avanzada con un costo estimado en un 60 % inferior al de modelos occidentales equivalentes no constituye un mero dato económico: representa un terremoto geopolítico que socava la doctrina estratégica basada en la acumulación cuantitativa de superioridad tecnológica.** Este modelo de eficiencia disruptiva — donde la excelencia algorítmica se combina con accesibilidad masiva — invalida los tradicionales cálculos de poder que han dominado ámbitos críticos como la ciberseguridad de infraestructuras jurídicas, la custodia de datos sensibles transfronterizos o las operaciones de guerra informática.

Esta transformación opera mediante un doble mecanismo paradójico. Por un lado, la naturaleza abierta del código de DeepSeek-R1 —al facilitar auditorías independientes y adopción global— promueve un ecosistema tecnológico más transparente y colaborativo. Países con recursos limitados pueden implementar herramientas de élite sin depender de actores extranjeros, fortaleciendo su soberanía digital. Sin embargo, esta misma apertura genera vulnerabilidades sistémicas: al exponer la arquitectura interna del sistema, se crean vectores de ataque potenciales que actores hostiles —desde cibercriminales hasta Estados revisionistas— podrían explotar para infiltrar sistemas jurídicos, manipular procesos legales automatizados o sustraer información clasificada. La ironía resultante es profunda: la transparencia que fortalece la confianza global simultáneamente debilita las barreras defensivas tradicionales.

Este dilema redefine, radicalmente, el concepto de soberanía en la era algorítmica. Los Estados ya no compiten solo mediante arsenales de hardware o reservas de datos, sino a través de su capacidad para gestionar la paradoja inherente a sistemas abiertos pero vulnerables. La ventaja estratégica se desplaza hacia quienes logren equilibrar innovación acelerada con marcos legales ágiles —normativas capaces de evolucionar al ritmo exponencial de la IA—, protección de infraestructuras críticas y preservación de derechos fundamentales en entornos digitales permeables. China, al posicionar a DeepSeek-R1 como estándar abierto, no solo exporta tecnología: impone un nuevo juego de reglas donde la seguridad nacional depende de la habilidad para navegar contradicciones entre transparencia y protección, entre colaboración global y autopreservación estratégica.

La implacable lógica de este paradigma obliga a reimaginar los pilares del derecho internacional y las políticas de defensa. ¿Cómo legislar sobre vulnerabilidades en código abierto utilizado simultáneamente por aliados y adversarios? ¿Qué mecanismos de responsabilidad aplicar cuando fallos en sistemas accesibles globalmente comprometen infraestructuras críticas en múltiples jurisdicciones? DeepSeek-R1, en su aparente neutralidad técnica, expone la obsolescencia de los marcos legales actuales y anuncia una nueva era donde la seguridad nacional será inseparable de la gobernanza algorítmica colaborativa —un desafío que exigirá tanto ingenio diplomático como innovación jurídica a escala planetaria—.

4.- Código Abierto vs Cerrado: La Transparencia en la Balanza

Hasta este punto, nuestra exploración de los Modelos de Lenguaje de Gran Escala (LLMs) se ha centrado, primordialmente, en sistemas de código cerrado o "Closed Source", como GPT-4 de OpenAI y Claude de Anthropic. Estos modelos, desarrollados por entidades privadas, se caracterizan por la opacidad de su arquitectura interna y la restricción de acceso a su código fuente, a menudo acompañada de un modelo de negocio basado en el pago por uso.

Sin embargo, para obtener una comprensión integral del panorama actual de los LLMs y sus implicaciones para la administración de justicia, es imperativo que también dirijamos nuestra atención hacia la contraparte de estos sistemas: los modelos de lenguaje de código abierto u "*Open Source*".

La distinción fundamental entre los Modelos de Lenguaje de Gran Escala (LLMs) de código abierto y cerrado radica en la accesibilidad y transparencia de su arquitectura, código fuente y datos de entrenamiento¹¹⁷. Esta diferencia, aparentemente técnica, tiene profundas implicaciones para su potencial aplicación en el ámbito de la administración de justicia.

Los LLMs de código cerrado, como GPT-4 y Claude, son desarrollados por entidades privadas y se caracterizan por la opacidad de su funcionamiento interno. El acceso a su código fuente y arquitectura está restringido y los detalles de sus datos de entrenamiento y algoritmos

¹¹⁷ Ilya Sutskever, "Open-Source vs. Closed-Source AI," en Inside OpenAI [Entire Talk], entrevistado por Ravi Belani, video, 4 min, Stanford eCorner, 26 de abril de 2023, <https://ecorner.stanford.edu/clips/open-source-vs-closed-source-ai/>

subyacentes a menudo se mantienen como secretos comerciales. Esta falta de transparencia puede plantear desafíos significativos cuando se considera su uso en un dominio tan sensible y de alta exigencia como el de marras.

La opacidad de los LLMs de código cerrado puede dificultar la evaluación de la equidad, imparcialidad y legalidad de las decisiones tomadas o asistidas por estos sistemas. Si el razonamiento y los procesos internos de un modelo no pueden ser examinados y comprendidos plenamente, se plantean preguntas sobre la legitimidad de su uso en un contexto donde la transparencia y el debido proceso son esenciales.

Al considerar la aplicación de LLM en el contexto legal y judicial, muchos sostienen que solo los modelos de código abierto pueden satisfacer los rigurosos estándares éticos y normativos necesarios para los SAD que potencialmente afecten los derechos y las libertades fundamentales de los ciudadanos. Entre las ofertas actuales de código abierto, dos de los contendientes más prometedores son Llama de Meta AI y la de Mixtral recientemente lanzada por Mistral AI Labs, laboratorio francés.

3.1.- Llama de Meta AI

El 18 de abril de 2024, Meta AI dio a conocer Llama 3¹¹⁸, la última iteración de su modelo de lenguaje a gran escala de código abierto que se está convirtiendo rápidamente en el nuevo estándar de referencia para LLM permisivamente licenciados. Con un modelo de pre-entrenamiento aclamado que supera a sus antecesores Llama 2 y LLaMA original, así como a los modelos propios de Meta de parámetros comparables, Llama 3 promete un rendimiento impresionante en una amplia gama de benchmarks y capacidades de IA generativa del lenguaje natural.

Fundamentalmente, Llama 3 se basa en modelos a pequeña escala de 8 mil millones y 70 mil millones de parámetros (denotados como 8B y 70B, respectivamente). Un parámetro en un modelo de inteligencia artificial es una variable que el modelo ajusta durante su entrenamiento para hacer predicciones precisas. Estos tamaños de modelo ofrecen un equilibrio atractivo entre

¹¹⁸ Meta. "Presentamos Meta Llama 3: modelo de lenguaje a gran escala más potente hasta la fecha." Publicado el 18 de abril de 2024. <https://www.meta.com/news/meta-llama-3>

capacidades de vanguardia y eficiencia computacional, ya que modelos más grandes generalmente pueden manejar tareas más complejas, pero requieren más recursos para funcionar.

El tokenizador de Llama 3 es una herramienta que transforma el texto en una serie de tokens, que son pequeñas unidades de significado como palabras o fragmentos de palabras. Este proceso permite que el modelo maneje y procese el texto de manera más eficiente, reduciendo el tiempo necesario para analizar y comprender grandes cantidades de información.

La atención de consulta grupal optimizada (Group Query Attention, o GQA) es una técnica avanzada que mejora la forma en que el modelo selecciona y se enfoca en la información más relevante dentro de un conjunto de datos. Al utilizar GQA, el modelo puede identificar y priorizar rápidamente los datos más importantes, lo cual mejora la velocidad de las inferencias, es decir, las predicciones o respuestas generadas por el modelo.

En conjunto, el tokenizador eficiente y la atención de consulta grupal optimizada permiten a Llama 3 realizar sus tareas de manera mucho más rápida que versiones anteriores del modelo. Esto se traduce en un procesamiento más ágil y en la capacidad de ofrecer respuestas más veloces y precisas.

Más allá de ser simplemente puntos de referencia, los modelos de Llama 3 superan a sus predecesores y competidores en varios casos de uso del mundo real, tales como la respuesta a preguntas, la generación de código, tareas de razonamiento y escritura creativa. Los modelos ajustados a las instrucciones basados en Llama 3 demuestran una mejorada "dirigibilidad" (la capacidad del modelo para seguir instrucciones de manera precisa y coherente), una alineación más fuerte (la capacidad del modelo para producir respuestas que estén en concordancia con valores humanos y expectativas) y una diversidad de respuesta ampliada (la capacidad del modelo para generar una variedad más amplia de respuestas), en gran parte, debido a técnicas de afinación innovadoras que aprovechan la optimización de políticas de recompensa (un proceso en el cual el modelo es entrenado utilizando recompensas para mejorar su desempeño en tareas específicas).

Desde una perspectiva de seguridad y responsabilidad, Meta también ha desarrollado una serie de herramientas y salvaguardas complementarias como Llama Guard 2, Code Shield y CyberSec Eval 2 para detectar y filtrar contenido dañino, prevenir la generación de código inseguro

y mitigar otros riesgos potenciales asociados con el uso indebido de Llama. El equipo también ha publicado una Guía de uso responsable (RUG)¹¹⁹ integral para ayudar a los desarrolladores externos a implementar los modelos de Llama de una manera ética y socialmente beneficiosa.

Sin embargo, Llama no está exenta de limitaciones y posibles inconvenientes. En primer lugar, a pesar de sus impresionantes capacidades multilingües que abarcan más de 30 idiomas, actualmente no se espera que Llama 3 alcance el mismo nivel de rendimiento en otros idiomas distintos del inglés. En segundo lugar, los modelos subyacentes aún se basan, en gran medida, en el aprendizaje supervisado a partir de conjuntos de datos de entrenamiento potencialmente sesgados, lo que plantea dudas sobre su equidad y objetividad.

Además, como muchos LLM, Llama puede ser propensa a "alucinar" o generar información factualmente incorrecta con alto grado de confianza, especialmente cuando se le hace una consulta sobre temas muy oscuros. Meta ha intentado mitigar parcialmente este problema mediante avisos de usuario contextuales, pero sigue siendo un riesgo inherente.

3.2.- Mixtral de Mistral AI: Mixture of Experts

Lanzado en diciembre de 2023 por Mistral AI, una nueva compañía de inteligencia artificial centrada en modelos de vanguardia, Mixtral representa un enfoque arquitectónico fundamentalmente diferente para construir LLM de alto rendimiento y eficientes en recursos. A diferencia de los modelos densos tradicionales como Llama y GPT-3 (porque, aunque se ha intentado mantener confidencial, se ha filtrado que GPT-4 utiliza la misma arquitectura¹²⁰) que aplican los mismos conjuntos de parámetros a cada token de entrada, Mixtral emplea una técnica conocida como mezcla dispersa de redes expertas o Mezcla de Expertos (SMoE).

Lanzado en diciembre de 2023 por Mistral AI, una nueva compañía de inteligencia artificial centrada en modelos de vanguardia, Mixtral representa un enfoque arquitectónico fundamentalmente diferente para construir modelos de lenguaje grande (LLM) de alto rendimiento

¹¹⁹ Meta Llama. "Responsible Use Guide: Your Resource for Building Responsibly.". <https://www.meta.com/responsible-use-guide>

¹²⁰ Ines Almeida, "Is GPT-4 a Mixture of Experts Model? Exploring MoE Architectures for Language Models," AI Insights (blog), 17 de Agosto de 2023, <https://www.nownextlaterai.com/insights/gpt-4-moe>

y eficientes en recursos¹²¹. A diferencia de los modelos densos tradicionales como Llama y GPT, que aplican los mismos conjuntos de parámetros a cada token de entrada (un token es una unidad de texto, como una palabra o un fragmento de una palabra), Mixtral emplea una técnica conocida como mezcla dispersa de redes expertas (SMoE, por sus siglas en inglés).

En un modelo Mixtral, el componente de feedforward (la parte del modelo encargada de procesar y transformar las entradas) en cada capa selecciona entre ocho grupos distintos de parámetros o "expertos". Para cada token en una secuencia dada, una red de enrutamiento (un sistema que decide qué expertos utilizar) elige los dos expertos más relevantes para procesar ese token específico y luego combina sus salidas de manera aditiva (sumando las salidas de los expertos seleccionados). Este enfoque permite que el modelo aproveche un conjunto masivo de parámetros (por ejemplo, 46.7 mil millones en total), mientras solo se invoca una fracción de ellos (por ejemplo, 12.9 mil millones) por token, lo que resulta en una inferencia (el proceso de hacer predicciones o generar texto) mucho más rápida y rentable.

De hecho, el logro más impresionante de Mixtral es que supera a Llama 2 70B en la mayoría de los puntos de referencia manteniendo una latencia y un costo de inferencia seis veces menores, lo que lo convierte en el mejor modelo general en términos de compensaciones costo/rendimiento. En particular, Mixtral coincide o supera el rendimiento de GPT3.5 en la mayoría de los puntos de referencia estándar, a pesar de tener potencialmente órdenes de magnitud menos parámetros y datos de entrenamiento

¹²¹ Mistral AI Team. "Mixtral of Experts: A High Quality Sparse Mixture-of-Experts." 11 de diciembre de 2023, <https://mistral.ai/news/mixtral-of-experts/>

	LLaMA 2 70B	GPT - 3.5	Mixtral 8x7B
MMLU (MCQ in 57 subjects)	69.9%	70.0%	70.6%
HellaSwag (10-shot)	87.1%	85.5%	86.7%
ARC Challenge (25-shot)	85.1%	85.2%	85.8%
WinoGrande (5-shot)	83.2%	81.6%	81.2%
MBPP (pass@1)	49.8%	52.2%	60.7%
GSM-8K (5-shot)	53.6%	57.1%	58.4%
MT Bench (for Instruct Models)	6.86	8.32	8.30

FIGURA 15: Rendimiento comparativo de Mixtral 8x7B frente a LLaMA 2 70B y GPT-3.5 en varios benchmarks¹²².

Esta tabla compara el rendimiento de tres modelos de inteligencia artificial: LLaMA 2 70B, GPT-3.5 y Mixtral 8x7B, en diversas tareas de evaluación. Los benchmarks incluyen:

- MMLU (Massive Multitask Language Understanding): este benchmark evalúa la capacidad del modelo para responder preguntas de opción múltiple en 57 temas diferentes, evaluando la comprensión general del lenguaje y el conocimiento de dominio específico.
- HellaSwag: una prueba de selección de opciones donde el modelo debe elegir la continuación más lógica de una historia o secuencia de eventos a partir de varias opciones, midiendo la coherencia y el sentido común.
- ARC Challenge (AI2 Reasoning Challenge): este benchmark desafía a los modelos con preguntas de opción múltiple diseñadas para evaluar el razonamiento y el conocimiento científico a nivel de educación básica y secundaria.

¹²² Ibid.

- WinoGrande: un benchmark de desambiguación pronombre-coreferencia donde el modelo debe resolver problemas de coreferencia pronombre en contextos complejos, evaluando la comprensión del texto y el razonamiento.
- MBPP (Mostly Basic Python Programming): esta prueba evalúa la capacidad del modelo para resolver problemas básicos de programación en Python, midiendo la comprensión de la lógica de programación y la sintaxis.
- GSM-8K (Grade School Math 8K): un conjunto de problemas de matemáticas a nivel de escuela primaria que evalúa la capacidad del modelo para realizar razonamientos matemáticos y resolver problemas aritméticos.
- MT Bench (Multi-Task Benchmark): este benchmark evalúa la capacidad del modelo para seguir instrucciones detalladas en múltiples tareas, reflejando su utilidad en aplicaciones prácticas de instrucción.

Los resultados muestran que Mixtral 8x7B supera a los otros dos modelos en la mayoría de los casos, destacando especialmente en MBPP con un 60.7 % y en MMLU con un 70.6 %.

Además de su excelente capacidad de inferencia, Mixtral se destaca en varios aspectos notables. Primero, puede manejar con gracia un contexto extremadamente largo de hasta 32k tokens, lo cual permite un razonamiento mucho más sofisticado y basado en el contexto. En segundo lugar, actualmente admite cinco idiomas principales (inglés, francés, italiano, alemán y español) con planes para expandirse a más en el futuro.

Quizás lo más intrigante es que los modelos Mixtral preentrenados se pueden ajustar a las instrucciones directas utilizando técnicas de afinación como el ajuste fino supervisado (*Supervised Fine-Tuning, SFT*) y la optimización de preferencias directas (*Direct Preference Optimization, DPO*). Estas técnicas permiten que los modelos sean más precisos y efectivos en tareas específicas.

El ajuste fino supervisado (SFT) es un proceso en el que un modelo preentrenado se entrena adicionalmente utilizando un conjunto de datos etiquetados. Esto permite que el modelo aprenda de ejemplos específicos y mejore su rendimiento en tareas relacionadas¹²³.

¹²³ Stephen M. Walker II, "What is supervised fine-tuning?", Klu.AI blog, <https://www.klu.ai/what-is-supervised-fine-tuning>

La optimización de preferencias directas (DPO) es una técnica en la que el modelo se entrena para optimizar directamente las preferencias de los usuarios o las métricas de rendimiento. Esto se hace mediante la retroalimentación directa y ajustando los parámetros del modelo para alinearse mejor con las preferencias deseadas¹²⁴.

El modelo Mixtral 8x7B Instruct resultante alcanza una puntuación impresionante de 8.3 en el benchmark MT-Bench. En este caso, como explicó supra, el benchmark MT-Bench mide la capacidad del modelo para seguir instrucciones.

Con esta puntuación, el modelo Mixtral 8x7B Instruct se convierte en el mejor modelo de código abierto con capacidades de seguimiento de instrucciones, mostrando un rendimiento comparable al de GPT-3.5. Esto significa que el modelo Mixtral puede realizar tareas complejas de manera similar a como lo hace GPT-3.5, pero con la ventaja de ser un modelo de código abierto, lo cual permite su uso y modificación por parte de la comunidad.

Desde una perspectiva ética, Mixtral parece presentar menos sesgos en el benchmark de consultas sesgadas (BQQ) en comparación con Llama 2. En general, los modelos Mixtral muestran sentimientos más positivos que Llama 2 en las dimensiones de lenguaje osado (BOLD), con variaciones similares dentro de cada dimensión. Esto sugiere que Mixtral puede ser intrínsecamente menos propenso a perpetuar estereotipos dañinos o generar contenido inapropiado.

¹²⁴ Rafael Rafailov et al., "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," arXiv, 13 de diciembre de 2023, Stanford University, <http://arxiv.org/abs/2305.18290>

	Llama 2 70B	Mixtral 8x7B
BBQ (higher is better)	51.50%	55.98%
BOLD (std) (lower is better)	0.094	0.084
gender	0.073	0.045
profession	0.073	0.087
religious_ideology	0.133	0.089
political_ideology	0.140	0.146
race	0.049	0.052

FIGURA 16: Comparación del rendimiento de los modelos Llama 2 70B y Mixtral 8x7B en las métricas *BBQ* y *BOLD*¹²⁵.

A pesar de estos puntos fuertes, Mixtral también tiene algunas desventajas en comparación con Llama. Primero, aunque está disponible públicamente, no tiene la misma escala de ecosistema o adopción que Llama ha cultivado con sus integraciones con plataformas líderes. En segundo lugar, Mixtral es un proyecto mucho más nuevo e inmaduro, sin el historial o la reputación establecidos de Meta AI.

3.3.- El Dilema de la IA Legal: entre la Capacidad y la Explicabilidad

La irrupción de los Large Language Models (LLMs) en el ámbito jurídico representa un hito histórico cuya trascendencia radica no solo en su sofisticación técnica, sino en su capacidad para reestructurar los fundamentos epistemológicos y operativos del derecho contemporáneo. Estos sistemas, desde modelos como GPT-4 hasta arquitecturas más recientes como Claude 3 Opus o Gemini 1.5, encarnan una convergencia sin precedentes entre avances en inteligencia artificial y las necesidades estructurales de sistemas legales cada vez más complejos. La escalada exponencial en parámetros —que superan los billones de operaciones— y la expansión de ventanas de contexto —capaces de procesar millones de tokens— han permitido a estas herramientas trascender el mero

¹²⁵ Ibid.

análisis estadístico, aproximándose a formas incipientes de comprensión contextual. Este progreso se sustenta en innovaciones técnicas como el Reinforcement Learning from Human Feedback (RLHF), que refina la alineación ética de los modelos y mecanismos de atención escalables, que optimizan la identificación de patrones en macrocorpus jurídicos. El resultado es una generación de sistemas que no solo replican, sino que en ciertos escenarios superan capacidades humanas en tareas como síntesis normativa, detección de inconsistencias legales o incluso resolución de exámenes profesionales, retos que hace un lustro se consideraban inalcanzables para la inteligencia artificial.

Sin embargo, la verdadera disruptión de estos modelos yace en su sinergia con las demandas intrínsecas de los sistemas jurídicos modernos. La habilidad de procesar y contextualizar vastos volúmenes de información legal —desde jurisprudencia histórica hasta legislación transnacional— resuelve uno de los dilemas más persistentes en la práctica legal: la gestión eficiente de datos dispersos y fragmentados. Modelos como Claude 3 Opus, al operar con ventanas de contexto expandidas, permiten localizar "agujas fácticas" en "pajares documentales", agilizando procesos críticos como la investigación precedencial o el contraste probatorio. Esta capacidad se ve potenciada por técnicas como el *chain-of-thought*, implementado en arquitecturas como OpenAI O1 o DeepSeek R1, que introduce un paradigma de transparencia algorítmica al exigir a la IA desplegar su razonamiento paso a paso, análogamente a como un juez fundamenta una sentencia. Tal trazabilidad mitiga la opacidad epistemológica de los sistemas de *caja negra*, facilitando su auditoría y alineación con principios procesales como la publicidad y la contradicción.

El caso de DeepSeek-R1, con su combinación de alto rendimiento, apertura algorítmica y costos reducidos, encapsula esta tensión dialéctica entre capacidad técnica y soberanía digital. Su emergencia no solo desafía la hegemonía tecnológica occidental, sino que redefine los términos del debate sobre gobernanza algorítmica: al democratizar el acceso a IA de élite, impone una reevaluación urgente de los marcos normativos internacionales, particularmente en lo concerniente a responsabilidad por decisiones automatizadas, protección de datos sensibles y equilibrio entre innovación y control democrático.

En este contexto, la implementación de LLMs en la administración de justicia exige un equilibrio delicado. Sistemas como Gemini 1.5 o Claude 3 Opus, capaces de localizar "agujas" fácticas en "pajares" documentales de millones de tokens, podrían revolucionar la investigación

legal y la redacción de sentencias. No obstante, su adopción requiere salvaguardias rigurosas: protocolos de validación humana, mecanismos de explicabilidad reforzada y auditorías continuas para detectar sesgos estructurales. La integración de técnicas de deliberative alignment y reinforcement learning from human feedback (RLHF) en modelos como O1 sugiere caminos promisorios para alinear estas herramientas con principios ético-jurídicos, aunque su eficacia última dependerá de la voluntad política para priorizar estándares públicos sobre intereses corporativos.

El futuro inmediato exige un doble movimiento estratégico. Por un lado, la adopción pragmática de estas herramientas para optimizar la administración de justicia —automatizando tareas rutinarias, identificando patrones jurisprudenciales ocultos, y asistiendo en la redacción de resoluciones complejas—. Por otro, el desarrollo urgente de marcos regulatorios adaptativos que, sin obstruir la innovación, salvaguarden principios democráticos esenciales. Esto implica protocolos estrictos de validación forense para sistemas de IA jurídica, auditorías continuas de sesgos y, sobre todo, el mantenimiento de la primacía humana en decisiones que afecten derechos fundamentales. La paradoja final radica en que, para preservar los valores centrales del Estado de Derecho en la era algorítmica, debemos ser tan innovadores en la gobernanza ética de la IA como lo son sus creadores en el avance técnico.

5.- La Vigencia Efímera del Estado del Arte Tecnológico: Una Guía Prospectiva

La incursión detallada en las arquitecturas y capacidades de los modelos de lenguaje de gran escala (LLMs) más prominentes —desde las generaciones fundacionales como GPT-3 hasta los sistemas avanzados de razonamiento encarnados en O1 y DeepSeek-R1— constituye un ejercicio analítico necesario, pero inherentemente **efímero**. Como se argumentó extensamente en el preámbulo de este capítulo, la inteligencia artificial no avanza de manera lineal, sino que se rige por una dinámica de **aceleración exponencial**, donde los ciclos de innovación se comprimen y las fronteras del "estado del arte" se redefinen con una celeridad sin precedentes.

Esta **volatilidad intrínseca** implica, sin ambages, que la fotografía técnica aquí presentada —identificando qué modelo específico ostenta la primacía en determinadas métricas o capacidades— poseerá una vigencia limitada. La propia secuencia histórica explorada, que evidencia el salto cuántico desde los modelos de 2020 hasta las proyecciones para 2025, es la prueba más palmaria de esta transitoriedad. Es más que plausible, es casi una certeza, que en el

lapso que medie entre la redacción de estas líneas y su lectura futura (incluso en un plazo tan breve como seis meses después), nuevos paradigmas algorítmicos hayan emergido, eclipsando las capacidades aquí descritas y reconfigurando el horizonte de lo tecnológicamente posible.

Sin embargo, reconocer esta caducidad no invalida la **función estructural** que cumple la exposición técnica precedente dentro del armazón argumentativo de esta tesis. Lejos de ser un mero apéndice descriptivo o una concesión a la curiosidad tecnológica, este análisis proporciona el **sustrato conceptual y empírico indispensable** para comprender la naturaleza, el alcance y, sobre todo, los *riesgos* que la inteligencia artificial plantea específicamente para la administración de justicia. Su valor reside no tanto en la perdurabilidad de los ejemplos concretos, sino en su capacidad para:

1. **Ilustrar los Paradigmas Tecnológicos Relevantes:** Al describir arquitecturas como los Transformers, los modelos basados en razonamiento (*chain-of-thought*) o las arquitecturas de "mezcla de expertos" (MoE), se dota al lector de una comprensión funcional de *cómo* operan los tipos de IA que actualmente se perfilan como candidatos para aplicaciones judiciales (análisis de texto legal, asistencia en la redacción, predicción basada en patrones, etc.).
2. **Identificar los Desafíos Técnicos con Implicaciones Jurídico-Éticas:** La discusión sobre las ventanas de contexto, la "opacidad" de las redes neuronales profundas (el problema de la "caja negra"), la propensión a "alucinaciones" o la generación de sesgos algorítmicos no son meros detalles técnicos, sino que señalan los **puntos neurálgicos** donde la tecnología colisiona con principios como la transparencia procesal, el derecho a la explicación, la igualdad ante la ley y la fiabilidad probatoria. Comprender estas limitaciones técnicas es esencial para valorar la idoneidad y los riesgos de implementar estas herramientas en un entorno judicial.
3. **Dimensionar la Velocidad del Cambio:** La propia exposición de la rápida sucesión de modelos (GPT-3 -> GPT-4 -> O1/DeepSeek) sirve para **subrayar la urgencia y la complejidad de la tarea regulatoria**. Demuestra por qué un enfoque normativo estático es inviable y por qué se requiere un marco adaptable y basado en principios perdurables, como el que intenta construir la Unión Europea.
4. **Ofrecer un Anclaje Concreto al Análisis Normativo Posterior:** La regulación no opera en el vacío. Al analizar el AI Act europeo o al proponer reformas para Costa Rica, es

fundamental tener en mente *qué tipo* de sistemas se están regulando. La discusión técnica previa permite entender *por qué* la UE clasifica la IA judicial como de "alto riesgo", *por qué* exige supervisión humana o *por qué* insiste en la calidad de los datos de entrenamiento. La técnica informa y justifica la norma.

En definitiva, la exposición técnica precedente, aunque volátil en sus detalles específicos, posee un **valor heurístico y contextual** irrenunciable para el análisis normativo que constituye el núcleo de esta tesis.

Ante esta realidad, y con el fin de ofrecer un **anclaje duradero** más allá de la instantánea tecnológica específica de este momento, se remite al lector interesado en la vanguardia técnica a plataformas dinámicas que monitorizan y comparan continuamente el rendimiento de los diversos modelos de IA disponibles. Estas herramientas sirven como una **guía prospectiva** para identificar qué sistemas se encuentran en la frontera del desarrollo y, por ende, cuáles podrían ser los candidatos más idóneos para futuras implementaciones que busquen incorporar lo más avanzado en la administración de justicia. Entre las plataformas de referencia más reconocidas se encuentran:

1. **Chatbot Arena (LMSYS Org):** Esta plataforma ofrece un *ranking* de modelos de lenguaje basado en la preferencia directa de miles de usuarios anónimos. Mediante comparaciones ciegas ("blind side-by-side battles"), los usuarios votan por la respuesta que consideran mejor entre dos modelos diferentes. Los resultados se agregan para generar una puntuación Elo (un método de calificación comparativa usado originalmente en ajedrez), que refleja la percepción general de calidad y utilidad de cada modelo en tareas conversacionales y de generación de texto. Es una métrica útil para evaluar la "popularidad" y la habilidad percibida en interacciones naturales.
 - *Sitio web oficial de Chatbot Arena:* <https://chat.lmsys.org/>
2. **Scale LLM Leaderboard (Scale AI):** Esta es una tabla de clasificación que evalúa modelos de lenguaje (tanto de código abierto como cerrado) utilizando un conjunto estandarizado de métricas cuantitativas en diversas tareas y benchmarks académicos (conocidos como "evals"). Estos benchmarks suelen medir capacidades específicas como el razonamiento matemático, la comprensión lectora, la generación de código, el conocimiento general, etc. (ej., MMLU, GSM8K). Proporciona una visión más técnica y objetiva del rendimiento de los modelos en tareas específicas, útil para seleccionar la herramienta más adecuada según la necesidad concreta (ej., un modelo fuerte en

razonamiento lógico podría ser más apto para análisis legal que uno fuerte solo en creatividad).

- *Sitio web oficial de Scale LLM Leaderboard,:* <https://scale.com/llm-leaderboard>

Consultar periódicamente estas (u otras) plataformas permitirá al lector futuro —ya sea académico, legislador, juez o tecnólogo— mantenerse actualizado sobre qué modelos lideran la vanguardia en capacidades relevantes para el ámbito jurídico y judicial, superando la inevitable obsolescencia de la información técnica específica contenida en este capítulo.

No obstante, es crucial subrayar que, si bien los *modelos específicos* y sus *capacidades puntuales* evolucionarán rápidamente, los *principios éticos*, los *desafíos jurídicos* y las *necesidades de regulación y gobernanza* analizados en los capítulos subsiguientes de esta tesis conservan una **validez estructural y conceptual** mucho más perdurable. La tensión entre eficiencia y garantías, la necesidad de transparencia algorítmica, la salvaguarda de la independencia judicial y la prevención de sesgos discriminatorios seguirán siendo los ejes centrales del debate, independientemente de cuál sea el modelo de IA más potente en un momento dado.

6.- El Ecosistema Legaltech: Contextualizando la IA como Vértice de la Transformación Tecnológica en el Derecho

Antes de adentrarnos de manera específica en las complejidades inherentes a la implementación de la inteligencia artificial (IA) en el núcleo mismo de la función jurisdiccional, resulta metodológicamente indispensable situar este fenómeno dentro del marco conceptual más amplio del *Legaltech*. Este término, acuñado para designar el vasto y heterogéneo conjunto de tecnologías y software diseñado para proveer, optimizar o transformar los servicios legales, representa el ecosistema tecnológico general del cual la IA judicial es una manifestación particularmente avanzada y sensible¹²⁶.

El Legaltech no es un constructo monolítico ni se limita exclusivamente a la inteligencia artificial. Abarca un espectro diverso de herramientas que han ido permeando progresivamente la praxis jurídica, desde soluciones relativamente sencillas orientadas a la gestión de despachos

¹²⁶ Richard Susskind, *Tomorrow's Lawyers: An Introduction to Your Future*, 2nd ed. (Oxford: Oxford University Press, 2017), 45-58. Recuperado de: <https://pdfroom.com/books/tomorrows-lawyers-an-introduction-to-your-future/NpgpZJQe5jr/download>

(como software de facturación o administración de casos), pasando por plataformas de investigación jurídica que automatizan la búsqueda y el análisis de legislación y jurisprudencia, hasta aplicaciones más sofisticadas que emplean el Procesamiento del Lenguaje Natural (PLN) para la revisión automatizada de contratos, la realización de procesos de *e-discovery* (descubrimiento electrónico de pruebas) o la generación de documentos legales estandarizados. Su propósito fundamental ha sido, tradicionalmente, incrementar la eficiencia operativa, reducir costos, democratizar el acceso a la información legal y, en última instancia, optimizar la prestación de servicios por parte de abogados, firmas y departamentos legales.

En este contexto evolutivo, la inteligencia artificial se erige no como un sinónimo de Legaltech, sino como una de sus vertientes más disruptivas y con mayor potencial transformador. Los avances en aprendizaje automático (*machine learning*), redes neuronales profundas (*deep learning*) y, más recientemente, los modelos de lenguaje de gran escala (LLMs) que hemos analizado previamente, han dotado a las herramientas Legaltech de capacidades cognitivas inéditas. Ya no se trata únicamente de automatizar tareas rutinarias o de gestionar información de manera más eficiente, sino de asistir –y potencialmente influir– en procesos que tradicionalmente requerían un alto grado de discernimiento humano, como el análisis predictivo de resultados judiciales, la identificación de patrones complejos en grandes volúmenes de evidencia o la argumentación jurídica asistida¹²⁷.

Así, la IA se integra en el ecosistema Legaltech como un componente avanzado que potencia y redefine las soluciones existentes. Plataformas que antes ofrecían búsquedas jurisprudenciales basadas en palabras clave, ahora incorporan motores semánticos impulsados por IA capaces de comprender el contexto y la intención de la consulta. Herramientas de revisión contractual que se limitaban a identificar cláusulas estándar, ahora pueden detectar riesgos, inconsistencias o desviaciones respecto de las mejores prácticas gracias al entrenamiento con millones de documentos legales¹²⁸.

Sentado este marco conceptual, la distinción entre el Legaltech en sentido amplio y la IA aplicada específicamente al ámbito *judicial* (objeto de la siguiente sección) resulta crucial.

¹²⁷ Daniel Martin Katz, "Quantitative Legal Prediction—or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry," *Emory Law Journal* 62, no. 4 (2013): 909-966.

Recuperado de: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2187752

¹²⁸ Dana Remus y Frank Levy, "Can Robots Be Lawyers? Computers, Lawyers, and the Practice of Law," *Georgetown Journal of Legal Ethics* 30, no. 3 (2017): 501-558. <https://dx.doi.org/10.2139/ssrn.2701092>

Mientras que gran parte del ecosistema Legaltech se orienta a optimizar la labor de los abogados, las firmas y los departamentos legales *fuerá* de la sala del tribunal, la implementación de IA en la administración de justicia (como apoyo a la decisión, predicción de riesgos, o incluso automatización de fallos en casos simples) incide, directamente, en la función pública de impartir justicia, afectando principios constitucionales medulares como la independencia judicial, el debido proceso y la tutela judicial efectiva.

Esta contextualización es crucial porque permite comprender que la IA judicial no surge *ex nihilo*, sino que es la punta de lanza de una transformación digital más amplia que afecta a todo el sector legal. Sin embargo, precisamente por su impacto directo en la función jurisdiccional y en los derechos fundamentales de los justiciables, la IA judicial demanda un escrutinio ético y regulatorio mucho más intenso que otras aplicaciones de Legaltech. Precisamente, por ello, la siguiente sección se abocará a desentrañar el potencial específico y los desafíos inherentes que la inteligencia artificial plantea cuando traspasa el umbral de los despachos de abogados para adentrarse en el delicado terreno de la decisión judicial.

7.- La IA en el Ámbito Judicial: Potencial y Desafíos

En un Estado de Derecho, la administración de justicia desempeña un papel crucial como garante de la paz social y el orden público. Los tribunales, como órganos encargados de dirimir las controversias y velar por el respeto a los derechos fundamentales, constituyen un pilar esencial para la convivencia armónica y el desarrollo de una sociedad justa y equitativa. Sin embargo, en las últimas décadas, los sistemas judiciales de todo el mundo se han enfrentado a una serie de retos sin precedentes que ponen a prueba su capacidad para cumplir, eficazmente, con su elevada misión.

El incremento exponencial de la litigiosidad, fruto de la creciente complejidad de las relaciones sociales y económicas, ha propiciado una sobrecarga de trabajo en los órganos jurisdiccionales, con el consiguiente incremento de los tiempos de respuesta y el riesgo de menoscabo de la calidad de las resoluciones. Asimismo, la sofisticación de ciertas materias objeto de controversia, como las derivadas de los avances científicos y tecnológicos, exige de los juzgadores un conocimiento cada vez más especializado y una mayor capacidad de análisis para desentrañar las intrincadas cuestiones fácticas y jurídicas que se les plantean.

Ante este escenario, la sociedad reclama de la administración de justicia una mayor eficiencia y celeridad, sin que ello implique una merma de las garantías procesales ni de la calidad

de las decisiones. Es en este contexto, donde la Inteligencia Artificial (IA) emerge como una herramienta prometedora para coadyuvar a la mejora del sistema judicial, la capacidad de estos sistemas para procesar ingentes volúmenes de información, detectar patrones y correlaciones y realizar predicciones con un elevado grado de fiabilidad, los convierte en un valioso aliado para optimizar la gestión de los asuntos, reducir los tiempos de tramitación y auxiliar a los juzgadores en la toma de decisiones.

No obstante, la irrupción de estos sistemas en un ámbito tan sensible como la administración de justicia no puede abordarse de manera acrítica ni precipitada. Resulta imprescindible un análisis riguroso y ponderado de sus posibilidades y límites, así como de las implicaciones éticas y jurídicas que su utilización conlleva. Solo a través de un debate sereno y plural, que involucre a todos los actores concernidos, podremos sentar las bases para una implementación responsable y garantista de estas tecnologías, que redunde en un fortalecimiento del Estado de Derecho y en una justicia más accesible, eficiente y de calidad.

Nos encontramos ante una encrucijada histórica en la que la administración de justicia, como piedra angular del Estado de Derecho, está llamada a evolucionar de la mano de la IA para hacer frente a los desafíos del siglo XXI. Un reto apasionante que exige altura de miras, rigor técnico y un firme compromiso con los valores superiores que inspiran nuestro ordenamiento jurídico.

➤ Breves Notas sobre la Doctrina de Casos Fáciles y Complejos

La distinción entre casos fáciles y difíciles ha sido un tema recurrente en la teoría del derecho, con profundas implicaciones para la concepción de la función judicial y el papel de la discrecionalidad en la aplicación de las normas. Según la caracterización clásica de MacCormick, los casos fáciles serían aquellos en los que la solución jurídica se deriva de forma clara y unívoca de las reglas establecidas, mientras que los casos difíciles requerirían un esfuerzo interpretativo y argumentativo adicional por parte del juez para resolver las dudas o lagunas planteadas¹²⁹.

Esta distinción entraña con la visión positivista del derecho como un sistema completo y coherente de reglas, en el que la labor del juez se limitaría a un mero silogismo subsuntivo. Como apunta Hart, los casos fáciles o evidentes se manifiestan donde "*los términos generales parecen*

¹²⁹ Neil MacCormick, Legal Reasoning and Legal Theory, Oxford University Press, Oxford, 1978, p. 227.

no necesitar interpretación y donde el reconocimiento de ejemplos parece no problemático o ‘automático’, y son solo los familiares, que se repiten constantemente en contextos similares, donde hay un acuerdo general en los juicios sobre la aplicabilidad de los términos clasificatorios”¹³⁰. Desde esta óptica formalista, sería concebible la sustitución del juez por sistemas de inteligencia artificial en la resolución de estos supuestos rutinarios, en la medida en que no implicarían una labor propiamente valorativa o decisoria.

Sin embargo, esta concepción ha sido cuestionada por autores como Dworkin, que defiende la tesis de la "única respuesta correcta" incluso en los casos difíciles. Según este autor, el derecho no se agota en las reglas positivas, sino que incorpora también principios morales más abstractos que el juez debe ponderar para hallar la solución más justa: "*Los principios tienen una dimensión que falta en las normas: la dimensión del peso o importancia. Cuando los principios se interfieren (la política de protección a los consumidores de automóviles interfiere con los principios de libertad de contratación, por ejemplo), quien debe resolver el conflicto tiene que tener en cuenta el peso relativo de cada uno. En esto no puede haber, por cierto, una mediación exacta, y el juicio respecto de si un principio o directriz en particular es más importante que otro será con frecuencia motivo de controversia. Sin embargo, es parte esencial del concepto de principio el que tenga esta dimensión, que tenga sentido preguntar qué importancia o qué peso tiene*"¹³¹. Esta labor de ponderación sería ineludible incluso en los casos aparentemente simples, en la medida en que la aplicación de cualquier norma implica siempre una reconstrucción interpretativa del derecho a la luz de sus propósitos y valores subyacentes.

Más allá de este debate teórico, lo cierto es que la frontera entre casos fáciles y difíciles dista de ser nítida y categórica. Como señala Schauer, la facilidad o dificultad de un caso no depende tanto de propiedades intrínsecas de este, sino más bien de factores contextuales como la claridad de las normas aplicables, la disponibilidad de precedentes análogos, la simplicidad de los hechos o la previsibilidad de las consecuencias¹³². Así, un mismo supuesto podría ser considerado fácil o difícil en función del grado de indeterminación del derecho, la complejidad de las

¹³⁰ H.L.A. Hart, *The Concept of Law*, 2^a ed. (Oxford: Oxford University Press, 1961), 126-127, <https://annas-archive.org/md5/fff38067a2ea7d435f3a14f0e2d27d88>

¹³¹ Ronald Dworkin, *Taking Rights Seriously*, trad. Marta Guastavino (Londres: Gerald Duckworth & Co. Ltd., 1984), 2^a ed., diciembre 1989, ISBN 84-344-1508-9, <https://img.lpderecho.pe/wp-content/uploads/2021/09/Descargue-en-PDF-Los-derechos-en-serio-de-Ronald-Dworkin-LP.pdf>

¹³² Frederick Schauer, "The Role of the Text: Easy Cases," en *Methods of Constitutional Interpretation*, http://fs2.american.edu/dfagel/www/Class%20Readings/Schauer/Schauer%20Easy%20Cases%20Only_CleanedUp.pdf

circunstancias concurrentes o incluso la habilidad argumentativa de los operadores jurídicos implicados.

Algunos autores han propuesto criterios alternativos para distinguir los casos susceptibles de automatización de aquellos que requieren un juicio humano. Así, Re y Solow-Niederman contraponen la noción de "justicia equitativa" frente a la de "justicia codificada"¹³³.

Los autores definen la justicia equitativa como aquella que:

"(...) implica tanto la reflexión sobre los valores establecidos por el sistema legal como la aplicación razonada de esos valores en su contexto. La justicia equitativa es más visible en fallos judiciales discretos que están regidos por estándares y aplicados a hechos determinados a través de procedimientos individualizados. Pero incluso las decisiones ampliamente aplicables gobernadas por la ley positiva, como los casos que involucran interpretación estatutaria, a menudo plantean importantes oportunidades para el juicio discrecional. A diferencia del desarrollo de políticas o la elaboración de normas en contextos administrativos o legislativos, la justicia equitativa aspira a aplicar principios consistentes y está preparada para dejar de lado patrones generales en favor de circunstancias únicas. Ese poder discrecional requiere legitimación y viene con restricciones. En particular, la justicia equitativa generalmente lleva una obligación de proporcionar una explicación particularizada y específica del caso que conecte los principios legales, aplicados a través de un proceso legal, con los hechos particulares en cuestión. Debido a su naturaleza discrecional, contextual y dinámica, la justicia equitativa puede parecer totalmente incompatible con procesos algorítmicos automatizados. Por ejemplo, ¿puede un procedimiento de decisión preestablecido realmente incorporar una idea como la misericordia o desarrollar un equilibrio sensible a los hechos de factores de mitigación en un caso penal?"¹³⁴

En esencia, la justicia equitativa reconoce que la aplicación rígida de reglas legales puede, en ocasiones, conducir a resultados injustos. Por lo tanto, otorga un papel fundamental al juicio

¹³³ Richard M. Re y Alicia Solow-Niederman, "Developing Artificially Intelligent Justice," Stanford Technology Law Review 22 (2019): 242-289, https://law.stanford.edu/wp-content/uploads/2019/08/Re-Solow-Niederman_20190808.pdf

¹³⁴ Ibid, p. 252.

moral discrecional de los jueces, quienes están llamados a moderar la ley a través de la equidad cuando las circunstancias particulares de un caso así lo exijan.

Esta concepción de la justicia como un ejercicio casuístico de discernimiento ético encuentra su máxima expresión en el ámbito penal, donde "la 'justicia equitativa' o el juicio moral discrecional a menudo se considera primordial"¹³⁵. Conceptos como la "malicia premeditada" o la posibilidad de ejercer "misericordia" en la determinación de la pena reflejan la profunda confianza depositada en la sabiduría de los jueces para sopesar factores intangibles, como el remordimiento o las circunstancias atenuantes, que no pueden reducirse a meras fórmulas algorítmicas.

En contraste, la justicia codificada representa una visión más mecanicista y deshumanizada de la judicación. La definen de la siguiente forma:

"La justicia codificada se refiere a la aplicación rutinaria de procedimientos estandarizados a un conjunto de hechos. Con el tiempo, los jueces pueden aplicar estos procedimientos estandarizados, que constituyen un conjunto de reglas—o un "algoritmo legal" no computarizado—a un gran número de casos. Por lo tanto, la justicia codificada precede a la IA. Por ejemplo, mucho antes de los avances recientes en IA, la justicia codificada era visible en las pautas federales de sentencia, así como en muchas rúbricas administrativas y quasi-administrativas similares. En general, la justicia codificada aspira a establecer el conjunto total de variables legalmente relevantes por adelantado, mientras descarta otros hechos y circunstancias que se puedan descubrir en procedimientos individualizados. El objetivo básico de dicha estandarización es reducir el espacio para la discreción humana en la judicación, disminuyendo así las oportunidades de arbitrariedad, sesgo y desperdicio, al tiempo que aumenta la eficiencia, consistencia y transparencia. En resumen, la justicia codificada ve los vicios de la discreción, mientras que la justicia equitativa ve sus virtudes. Por lo tanto, la justicia codificada tiende a eliminar la necesidad de cualquier explicación, restricción o legitimación aparte de la adherencia a los procedimientos estandarizados en sí mismos. En otras palabras, el poder y la autoridad del juez se consideran no discretionales y derivan de cualquier entidad que haya creado el algoritmo legal relevante. Así que, cuando los sistemas algorítmicos analógicos o digitales reemplazan cualquiera de las funciones multifacéticas de los

¹³⁵ Ibid, p. 246.

tribunales de primera instancia y de apelación, el compromiso del poder judicial con la toma de decisiones discrecional y razonada se verá bajo presión"¹³⁶.

Los autores advierten que la creciente adopción de la IA de adjudicación, impulsada por su promesa de eficiencia y aparente imparcialidad, tenderá a favorecer la justicia codificada sobre la equitativa. Sin embargo, Re y Solow-Niederman no descartan por completo la posibilidad de que la IA pueda, en principio, incorporar elementos de justicia equitativa. Sugieren que "*la IA de adjudicación podría preservar o incluso fomentar la justicia equitativa*"¹³⁷ mediante la integración de "*cierta cantidad de toma de decisiones de IA junto con la reflexión y deliberación humanas*"¹³⁸. No obstante, reconocen que las limitaciones tecnológicas actuales, particularmente en lo que respecta a la interpretabilidad de los sistemas de aprendizaje automático profundo, hacen que "al menos en el futuro cercano, la IA de adjudicación no encarnará la justicia equitativa"¹³⁹.

Los autores destacan además que, incluso si la IA eventualmente se vuelve más interpretable, las fuerzas del mercado y los incentivos económicos de los desarrolladores privados de IA empujarán su desarrollo hacia la justicia codificada. Como señalan, "*los incentivos económicos [...] catalizarán aún más el impulso para desarrollar y desplegar programas de toma de decisiones basados en datos*"¹⁴⁰, favoreciendo la eficiencia y uniformidad sobre la discreción y el juicio humano.

En una línea similar, la noción de "procedimientos testigo" propuesta en el Proyecto de Ley de Eficiencia Procesal¹⁴¹ español de 2021 alude a aquellos casos que, "*compartiendo una identidad sustancial de objeto con otros procedimientos en curso, sirven de modelo para la resolución de estos últimos*" (art. 438 ter). Así los conceptualiza y justifica dicha propuesta legislativa:

"Por ello, en esta ley se busca dotar de nuevas herramientas a los órganos jurisdiccionales, así como a los justiciables, que permitan dar una respuesta adaptada,

¹³⁶ Ibid, p. 253.

¹³⁷ Ibid, p. 258.

¹³⁸ Ibid.

¹³⁹ Ibid, p. 260.

¹⁴⁰ Ibid, p. 270.

¹⁴¹ Boletín Oficial de las Cortes Generales. Congreso de los Diputados. XIV Legislatura. Serie A: Proyectos de Ley, 8 de junio de 2023. Informe de la Ponencia Núm. 97-4, 121/000097 Proyecto de Ley de medidas de eficiencia procesal del servicio público de Justicia. Recuperado de: https://www.congreso.es/public_oficiales/L14/CONG/BOCG/A/BOCG-14-A-97-4.PDF

eficaz y ágil a las pretensiones que se sustancien en el particular ámbito al que nos referimos. Una de las soluciones operadas por esta ley para la tramitación de este modo de litigar en masa es la incorporación del sistema de tramitación de los llamados «procedimientos testigo».

El procedimiento testigo es una vía que se articula para dar respuesta a demandas con identidad sustancial de objeto sin necesidad de tramitar todas ellas. Así, previa dación de cuenta por el letrado o letrada de la Administración de Justicia o a solicitud de la parte actora o demandada, se permite al juez o jueza elegir un procedimiento que se tramitará con carácter preferente, suspendiéndose el curso del resto de procedimientos en los que se dé aquella identidad. Una vez se dicte sentencia en el procedimiento testigo y adquiera firmeza, se requeriría a los afectados por los procedimientos suspendidos para que puedan solicitar la extensión de los efectos de la sentencia de referencia, continuar el procedimiento suspendido o desistir del mismo. De este modo se evita la tramitación simultánea o sucesiva de procedimientos judiciales sustancialmente idénticos en aras de garantizar un principio de economía procesal concebido de una manera mucho más amplia.

*Existen importantes razones para incorporar este sistema a nuestra regulación procesal civil en esta materia concreta ya que, en muchas ocasiones, los actores utilizan demandas o plantillas iguales o similares para el ejercicio de las mismas pretensiones, de modo que un universo muy amplio de perjudicados termina litigando con demandas prácticamente idénticas. De hecho, se ha generalizado un modo de litigación en masa en el que se utilizan plataformas informáticas no solo para captar clientes, sino también para la gestión de las demandas en las distintas fases. Teniendo en cuenta los extremos advertidos, es previsible que la regulación de este procedimiento testigo reducirá notablemente la litigación en masa, **en especial los procedimientos sobre nulidad por abusividad de las condiciones generales de la contratación en los que haya que valorar únicamente elementos objetivos**, y evitará la necesidad de completa tramitación de los procedimientos ya iniciados con identidad sustancial de objeto, lo que supondrá un alivio muy considerable en las cargas de trabajo de los órganos judiciales, reforzándose además la homogeneidad en las respuestas de la Justicia ante esta tipología de procedimientos. En relación a esta misma cuestión, y por exactamente los mismos motivos indicados para el procedimiento*

testigo, esta ley también regula el mecanismo procesal de extensión de efectos, importado también de la Ley reguladora de la Jurisdicción Contencioso-administrativa. Para la litigación en masa a la que se alude, la regulación actual de la extensión de efectos en acciones colectivas se ha mostrado claramente insuficiente. Como se ha dicho, los litigios en esta materia se han demostrado absolutamente repetitivos, y los eventuales obstáculos que puedan alegarse sobre la posible indefensión por falta de prueba chocan con la realidad de que, en la práctica totalidad de los procesos, no se pide otra que la documental

„142.

En este anteproyecto se proyecta que el litigante pueda solicitar la extensión de los efectos de la sentencia del procedimiento testigo cuando se cumplan las siguientes condiciones:

- Que la sentencia sea firme y definitiva.
- Que los interesados se encuentren en una situación jurídica idéntica a la de los beneficiados por el fallo.
- Que el demandado sea el mismo o su sucesor legal.
- Que no sea necesario realizar un control de transparencia de la cláusula ni evaluar posibles vicios en el consentimiento del contratante.
- Que las condiciones generales de contratación sean sustancialmente idénticas a las examinadas en la sentencia cuyos efectos se desean extender.

Así, se trataría de litigios altamente estandarizados en los que la solución dada a uno de ellos sería extrapolable al resto, facilitando así su tratamiento automatizado. Otro posible criterio distintivo sería el carácter preceptivo o facultativo de la defensa letrada, que revelaría indiciariamente la mayor o menor complejidad del asunto.

Sin embargo, estas aproximaciones conceptuales no están exentas de controversia, en la medida en que parecen sugerir una visión excesivamente mecanicista de la función judicial, reduciéndola a una mera labor subsuntiva de encaje del caso en el supuesto de hecho de la norma. Frente a ello, se alza una concepción más compleja y valorativa de la tarea del juez, como un agente llamado a realizar en cada caso el proyecto de justicia ínsito en el ordenamiento jurídico, ponderando las circunstancias concurrentes y los principios y valores en juego para hallar la solución más justa.

¹⁴² Ibid.

Desde esta perspectiva, el proceso judicial se configura como un escenario de diálogo y argumentación en el que el juez, asistido por las partes y los demás operadores jurídicos, ha de construir la respuesta más adecuada al caso, conjugando la previsibilidad de las normas con la equidad que reclama la justicia del caso concreto. Una visión que entraña con la clásica distinción aristotélica entre justicia legal y justicia natural y que sitúa al juez como un mediador prudencial entre la generalidad de la ley y la singularidad del litigio.

Esta tensión entre la concepción logicista y la concepción valorativa de la función judicial tiene importantes implicaciones para el uso de la IA en el proceso. Así, desde una óptica puramente subsuntiva, los casos fáciles serían aquellos en los que, por su simplicidad y reiteración, sería admisible la sustitución del juez por sistemas automatizados de decisión, basados en la detección de patrones y correlaciones en un ingente corpus de resoluciones previas. Por el contrario, los casos complejos requerirían indefectiblemente la intervención humana, ya sea para valorar las particularidades del supuesto, ponderar los principios en conflicto o modular la solución normativa para adecuarla a las exigencias de la justicia material.

Ahora bien, incluso desde una concepción más rica y compleja de la labor judicial, es posible identificar ciertos casos que, por su escasa entidad o nula controversia jurídica, admitirían un tratamiento automatizado sin merma de garantías. Tal sería el caso de los procesos monitorios, las reclamaciones de cantidad líquida y vencida o los juicios de faltas por hechos flagrantes y reconocidos, en los que la actuación del juez se limitaría a una constatación formal de los presupuestos legales para la estimación de la pretensión.

En estos supuestos, la utilización de LLMs que incorporen los criterios jurídicos de decisión, podría suponer una notable descarga de trabajo para los jueces, permitiéndoles concentrar sus esfuerzos en los asuntos de mayor enjundia. Eso sí, siempre que se adopten las debidas cautelas para garantizar la transparencia, trazabilidad y equidad de los algoritmos empleados, así como el derecho de los justiciables a una revisión humana de la decisión automatizada.

Más allá de estos casos puntuales, la utilidad de la IA en el ámbito judicial parece proyectarse fundamentalmente en una dirección auxiliar o de apoyo a la decisión humana. Así, en los procesos de mayor complejidad fáctica o jurídica, los sistemas de "justicia predictiva" basados en el análisis masivo de resoluciones previas pueden proporcionar al juez una valiosa orientación sobre el sentido probable del fallo, el rango de la pena o la indemnización esperable, o los

argumentos habitualmente empleados en supuestos análogos. Una información que, lejos de condicionar el criterio judicial, ha de concebirse como un input más en el proceso deliberativo que culmina con la sentencia, entendida como un acto de razón prudencial que aspira a realizar en el caso concreto los valores de justicia, equidad y proporcionalidad.

En definitiva, la distinción entre casos fáciles y complejos, pese a su indudable utilidad como criterio orientativo para discernir el encaje de la IA en el proceso, no puede erigirse en un dogma incontrovertible que propicie una cesión acrítica de la función judicial a la tecnología. Por el contrario, debe ser objeto de una reflexión constante y matizada que, partiendo de una concepción axiológica de la labor del juez, identifique aquellos espacios en los que la automatización puede reportar eficiencias sin comprometer la calidad de la tutela dispensada. Y, en todo caso, preservando siempre la centralidad de la persona como titular del derecho a una decisión justa y equitativa, que tome en consideración las circunstancias singulares de su caso.

➤ **Implementación de la IA en Decisiones Judiciales**

- **Sustitución en Casos Fáciles**

La aplicación de sistemas de inteligencia artificial, y en particular de los grandes modelos de lenguaje (LLMs), para la resolución automatizada de casos judiciales sencillos y repetitivos, se presenta como una de las opciones más prometedoras para aliviar la carga de trabajo de los tribunales y mejorar la eficiencia de la administración de justicia. Se trataría de aquellos supuestos en los que, por su escasa complejidad fáctica y jurídica, y por su alto grado de estandarización, la intervención humana parece prescindible, o incluso desaconsejable en aras de una mayor celeridad, coherencia y predictibilidad de las resoluciones.

En litigios como las reclamaciones de cantidad derivadas de impagos de facturas, las demandas de desahucio por falta de pago o los procedimientos monitorios, por citar solo algunos ejemplos, la labor del juez se limita en la mayoría de los casos a una mera constatación del cumplimiento de los requisitos legales y a la aplicación mecánica de las consecuencias previstas en la norma. Son procesos en los que apenas hay margen para la interpretación normativa, la valoración probatoria o la apreciación de las circunstancias específicas del caso, y en los que la respuesta judicial está prácticamente predeterminada por la concurrencia de unos presupuestos fácticos tipificados.

En estos supuestos, la utilización de LLMs entrenados con un gran volumen de resoluciones judiciales previas permitiría automatizar la decisión del caso mediante la identificación de los patrones fácticos y jurídicos relevantes y la generación de una propuesta de resolución adaptada a estos. El modelo, tras analizar la demanda y la documentación aportada, y contrastarla con su base de conocimiento jurisprudencial, podría determinar si se cumplen los requisitos legales para estimar la pretensión y, en su caso, generar automáticamente el correspondiente auto o sentencia, incorporando los fundamentos de derecho pertinentes y los pronunciamientos necesarios para su ejecución.

La principal ventaja de este enfoque radica en la potencial reducción de los tiempos de respuesta judicial en este tipo de litigios de alta frecuencia y baja complejidad. La automatización de las tareas más rutinarias y mecánicas permitiría incrementar, significativamente, la capacidad resolutiva de los tribunales, al tiempo que liberaría a los jueces de una carga de trabajo repetitiva para concentrarse en los asuntos más complejos y relevantes. Ello redundaría no solo en una mayor celeridad y eficiencia de la justicia, sino, también, en una mejor calidad y fundamentación de las resoluciones en aquellos casos que sí requieren de una reflexión jurídica más profunda y particularizada.

No obstante, la automatización de las decisiones judiciales, incluso en los casos más simples, no está exenta de riesgos y desafíos que es preciso afrontar. El primero y más evidente es el relativo a la transparencia y controlabilidad de los sistemas utilizados. El problema de la opacidad o "caja negra" de los grandes modelos de lenguaje (LLMs) ha sido uno de los principales desafíos para la investigación en inteligencia artificial. Estos modelos, a pesar de su impresionante capacidad para procesar y generar lenguaje natural de manera coherente y contextualizada, operan de una forma que resulta difícil de interpretar y auditar para los humanos. Esta falta de transparencia plantea serias preocupaciones en cuanto a la confiabilidad, equidad y alineación de valores de estos sistemas, especialmente cuando se contemplan aplicaciones de alto impacto social como la toma de decisiones judiciales.

Sin embargo, el horizonte parece estar iluminándose con los recientes avances en el campo de la interpretabilidad mecanicista de los LLMs. Un hito particularmente significativo en este camino fue la publicación, el 21 de mayo de 2024, del artículo *Scaling Monosemantics*:

*Extracting Interpretable Features from Claude 3 Sonnet*¹⁴³ por parte del equipo de interpretabilidad de Anthropic. Este trabajo representa un paso crucial hacia la apertura de la caja negra de los LLMs, al demostrar la extracción a gran escala de características interpretables del LLM de tamaño medio de Anthropic, Claude 3 Sonnet.

Para entender la importancia de este logro, primero debemos comprender un poco sobre cómo funcionan los LLMs. Estos modelos procesan el lenguaje transformándolo en complejos patrones numéricos llamados “espacios de activación”. La hipótesis es que, dentro de estos espacios, los conceptos significativos se codifican como direcciones específicas, conocidas como “características” o “features”.

Para ilustrar este concepto, podemos hacer una analogía con el proceso de aprendizaje humano. Cuando enseñamos a alguien a reconocer un concepto, como, por ejemplo, un perro, lo hacemos mostrándole una gran variedad de ejemplos que abarcan diferentes tamaños, formas, colores y razas. A medida que la persona observa estos ejemplos, su mente comienza a identificar los patrones comunes que definen a un perro, independientemente de sus variaciones individuales. En esencia, aprende a reconocer las “características” esenciales de un perro.

De manera similar, cuando un modelo de lenguaje es entrenado con una gran cantidad de texto, comienza a identificar patrones y conceptos recurrentes. Estos conceptos pueden ser tanto concretos, como personas, lugares u objetos específicos, como abstractos, como emociones, relaciones o tipos de eventos.

Sin embargo, a diferencia de los humanos, los modelos de lenguaje representan estos conceptos de una manera más abstracta y matemática. Cada concepto o “característica” se convierte en una dirección única en el “espacio de activación” del modelo, un patrón específico de actividad que se propaga a través de las neuronas artificiales del modelo. La complejidad surge cuando consideramos que, en los modelos de lenguaje tradicionales, estas características están altamente entrelazadas y superpuestas entre sí. Es como si todos los conceptos que el modelo ha aprendido estuvieran enredados en una compleja red de interconexiones, donde es difícil aislar un concepto individual del resto.

¹⁴³ Adly Templeton et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet," Anthropic (21 de mayo de 2024). <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>

Aquí es donde entra en juego el trabajo de Anthropic. Utilizando una técnica llamada "*aprendizaje de diccionario disperso*" (*sparse dictionary learning*), el equipo pudo extraer características interpretables del modelo Claude 3 Sonnet de Anthropic. El avance significativo logrado por el hallazgo derivado de esta nueva técnica fue, precisamente, la capacidad de desenredar algunos de estos conceptos. Pudieron identificar direcciones específicas en el espacio de activación que correspondían a conceptos únicos e interpretables, características individuales que podrían ser aisladas y estudiadas.

Lo que el equipo de Anthropic logró con su técnica fue, esencialmente, proyectar este espacio de alta dimensionalidad en un "diccionario" de características interpretables. Cada una de estas características representa una dirección específica en el espacio de activación del modelo que corresponde a un concepto único y coherente. Esta correspondencia uno a uno entre características y conceptos es lo que las hace interpretables: podemos entender qué concepto representa cada característica y estudiar cómo el modelo lo utiliza.

Pero el verdadero poder de este descubrimiento radica en que estas características no son meramente descriptivas, sino que tienen un impacto causal en el comportamiento del modelo. Los investigadores demostraron que, al manipular la actividad de una característica específica, es posible influir en las respuestas y decisiones del modelo de maneras predecibles y consistentes con el concepto que representa esa característica. Esta capacidad de intervención abre un mundo de posibilidades para auditar, controlar y guiar el comportamiento de estos sistemas.

Estos hallazgos tienen implicaciones profundas para el uso responsable y ético de los LLMs. Imaginemos un futuro cercano en el que se utiliza un LLM para ayudar en la toma de decisiones judiciales en casos simples y rutinarios. Antes del trabajo de Anthropic, no habría forma de saber qué conceptos o sesgos podrían influir en las decisiones del modelo. Pero ahora, con la capacidad de identificar y analizar características interpretables, podríamos auditar efectivamente el modelo en busca de sesgos problemáticos y quizás incluso corregirlos antes de ponerlo en uso.

Por supuesto, este trabajo es solo el comienzo. Quedan muchos desafíos por delante, desde mejorar la eficiencia de las técnicas de aprendizaje de diccionario hasta abordar cuestiones de superposición de características entre las capas del modelo. Pero el hecho de que estas características interpretables puedan ser descubiertas y manipuladas abre un mundo de posibilidades.

Otro desafío crucial es el relativo a la calidad y representatividad de los datos utilizados para el entrenamiento de los LLMs. Si los modelos son alimentados con un conjunto de resoluciones judiciales previas que adolecen de sesgos históricos o que no reflejan la diversidad social y cultural de los justiciables, existe el riesgo de que tales sesgos se perpetúen e incluso se amplifiquen en las decisiones automatizadas. Piénsese, por ejemplo, en el impacto discriminatorio que podría tener un LLM entrenado mayoritariamente con sentencias de desahucio dictadas en un contexto de crisis económica y que, por tanto, sobrerepresentan a determinados colectivos vulnerables.

Para conjurar este peligro, es necesario asegurar que los conjuntos de datos empleados para el desarrollo de LLMs judiciales sean lo más amplios, diversos y actualizados posible, incorporando resoluciones de distintos órganos jurisdiccionales, períodos temporales y perfil de justiciables. Asimismo, deben establecerse mecanismos de monitorización continua de los modelos para detectar posibles desviaciones o efectos discriminatorios, así como protocolos de retroalimentación que permitan afinarlos y corregirlos de manera dinámica a partir de la experiencia aplicativa.

Un ejemplo destacado de la aplicación de la IA para la sustitución del juez en casos sencillos son las denominadas "Smart Courts" de China¹⁴⁴, implementadas en los tribunales de Beijing y Guangzhou para la resolución automatizada de litigios menores en materia de comercio electrónico, propiedad intelectual o derecho del consumo. En agosto de 2017, se inauguró el Tribunal de Internet de Hangzhou en la ciudad de Hangzhou, provincia de Zhejiang, la cual es considerada la capital del comercio electrónico en China, ya que alberga la sede central de Alibaba. Este tribunal tiene la competencia para gestionar una variedad de casos relacionados con el internet, tales como disputas contractuales derivadas de compras y servicios en línea. A través de una plataforma web denominada 'Plataforma de Litigación del Tribunal de Internet de Hangzhou', todos los procedimientos judiciales pueden realizarse en línea, desde la presentación del caso y la notificación de documentos judiciales hasta el intercambio y examen de pruebas, la audiencia en línea y la emisión del fallo, aunque el tribunal puede optar por utilizar un proceso presencial para

¹⁴⁴ Changqing Shi, Tania Sourdin y Bin Li, "The Smart Court – A New Pathway to Justice in China?", International Journal for Court Administration 12, no. 1 (2021): p. 4-19, , <https://storage.googleapis.com/jnl-up-j-ijca-files/journals/1/articles/367/submission/proof/367-1-1754-2-10-20210311.pdf>

gestionar la audiencia. Un año después, se establecieron dos tribunales de Internet adicionales con plataformas de litigación en línea similares en Beijing y Guangzhou¹⁴⁵.

En este sentido, el juez Qian Du, presidente del Tribunal de Internet de Hangzhou, señaló en 2019 que, en sus dos años de funcionamiento, el tribunal había emitido alrededor de 20,000 sentencias y el tiempo promedio de audiencia para cada caso se había reducido en un 65 % en comparación con las audiencias presenciales¹⁴⁶.

Dentro de este sistema inteligente, los tribunales de Beijing emplean Smart Judge, un software que simula el proceso de pensamiento judicial. El software identifica las cuestiones legales presentadas por un caso, recupera materiales pertinentes para su resolución y recomienda una disposición. De manera similar, se está alentando a los jueces de Hainan, por parte del tribunal superior provincial, a adoptar un “sistema inteligente” que combina el procesamiento de lenguaje natural, gráficos de conocimiento y aprendizaje profundo para destilar la esencia de un caso y formular un juicio basado en decisiones anteriores. Elogiada por el Tribunal Popular Supremo de China como un modelo por seguir, se dice que esta práctica mejora la uniformidad de las sentencias y reduce a la mitad el tiempo necesario para emitir un fallo¹⁴⁷.

En Europa, países como Reino Unido, Francia o Estonia también han planteado iniciativas similares, si bien con un alcance más limitado y sujetas a mayores garantías procesales. Así, el gobierno británico llegó a proponer en 2017 la resolución automatizada de delitos leves flagrantes, como el hurto en tiendas o la posesión de cannabis, en aquellos casos en que el acusado se declarara culpable y aceptara someterse al procedimiento y restringido a los siguientes delitos no castigables con prisión: evasión de tarifa ferroviaria, evasión de tarifa de tranvía y posesión de caña y línea sin licencia¹⁴⁸.

¹⁴⁵ Ibid, p. 11.

¹⁴⁶ Ibid.

¹⁴⁷ Benjamín Minhao Chen y Zhiyu Li, "How Will Technology Change the Face of Chinese Justice?" Columbia Journal of Asian Law 34, no. 1 (2020): 18, <https://www.google.co.cr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjYpaXejAKGAXERDABHQ4wDgsQFnoECA4QAO&url=https%3A%2F%2Fjournals.library.columbia.edu%2Findex.php%2Fcjal%2Farticle%2Fdownload%2F7484%2F3923%2F14211&usg=AOvVaw3C4YIaB6W8Hd4WNo5BrfsZ&o pi=89978449>

¹⁴⁸ Lord Chancellor y Secretario de Estado de Justicia. “Transformando nuestro sistema de justicia: estrategia digital asistida, convicción automática en línea y pena estándar estatutaria, y composición de paneles en tribunales. Respuesta del gobierno”. Presentado al Parlamento por mandato de Su Majestad, febrero de 2017. Ministerio de Justicia Británico, p.8. [transforming-our-justice-system-government-response.pdf\(publishing.service.gov.uk\)](transforming-our-justice-system-government-response.pdf(publishing.service.gov.uk))

Aunque la propuesta fue finalmente descartada debido a las fuertes reticencias suscitadas en la abogacía y la judicatura¹⁴⁹, sentó un importante precedente sobre el potencial uso de la IA en la justicia penal. En cualquier caso, el Ministerio de Justicia británico ha seguido apostando por la digitalización de la administración de justicia, con iniciativas como el desarrollo de una plataforma de resolución de litigios de escasa cuantía mediante el programa denominado *HM Courts & Tribunals Service*.

Es un programa de 1,2 mil millones de libras que se ejecuta desde 2016 hasta 2023 y está diseñado, en parte, para incorporar las muy necesarias tecnologías digitales modernas en la infraestructura y los procesos de los servicios de tribunales y tribunales administrativos. El programa es ambicioso, la escala del cambio requerido es desalentadora y, como consecuencia, los plazos se han modificado recientemente para extender la duración del programa. Las reformas, una vez implementadas, tienen como objetivo introducir innovaciones digitales específicas. **Los litigantes podrán iniciar procedimientos en línea en los tribunales civiles y de familia. Habrá un proceso para rastrear las apelaciones en los tribunales administrativos. Muchas audiencias podrán realizarse mediante enlace de video, incluidas algunas solicitudes civiles y audiencias de prisión preventiva en los tribunales penales. Para algunos casos penales, será posible declararse culpable en línea y recibir una condena legal automática en línea por delitos como el impago de licencias de televisión.** Algunos sistemas, como la digitalización de la transmisión de información dentro de los tribunales, probablemente abarcarán todos los dominios¹⁵⁰.

Por su parte, Francia contempló en su plan de reforma judicial para el periodo 2018-2022 una propuesta de notable relieve que perseguía la implantación de una herramienta automatizada de resolución de conflictos civiles, con una competencia limitada a litigios no superiores a los 6.000 euros. No obstante, dicha propuesta ha sido suspendida por el momento.¹⁵¹ Por otra parte, Estonia se ha destacado como un país pionero en la materia, al proponer en 2019 la creación de un sistema predictivo de resolución de asuntos sencillos, cuya esfera de aplicación estaría acotada a

¹⁴⁹ The Law Society Commission on the Use of Algorithms in the Justice System, *Algorithms in the Criminal Justice System* (Londres: The Law Society of England and Wales, junio de 2019), <https://www.lawsociety.org.uk/topics/research/algorithm-use-in-the-criminal-justice-system-report>

¹⁵⁰ Ibid, p. 58.

¹⁵¹ Miguel de Asís Pulido, "La justicia predictiva: tres posibles usos en la práctica jurídica," en *Inteligencia Artificial y Filosofía del Derecho*, dir. Fernando H. Llano Alonso, coord. Joaquín Garrido Martín y Ramón Valdivia Jiménez (Murcia: Ediciones Laborum, 2022), 299.

litigios con una cuantía no superior a los 7.000 euros. En este último proyecto, la utilización de la resolución automatizada se encuentra especialmente sujeta a apelación¹⁵².

Más allá de estos ejemplos concretos, lo cierto es que el interés por la aplicación de la IA en la resolución de casos sencillos y repetitivos se ha extendido en los últimos años a numerosos países, tanto en Europa como en otras regiones del mundo. Así lo atestiguan iniciativas como el programa "Prometea" en Argentina, que utiliza técnicas de machine learning para predecir la solución de casos con un 96 % de precisión¹⁵³.

La aplicación de sistemas de inteligencia artificial y, particularmente, de los grandes modelos de lenguaje (LLMs), en la resolución automatizada de casos judiciales sencillos y repetitivos se vislumbra como un paso trascendental en la senda hacia una administración de justicia más eficiente, ágil y coherente. Esta innovación tecnológica, sabiamente implementada, puede convertirse en un valioso aliado de nuestros órganos jurisdiccionales, permitiéndoles focalizar sus esfuerzos y recursos en aquellos asuntos que, por su complejidad fáctica o jurídica, demandan un examen más exhaustivo y particularizado.

Sin embargo, sería un error sucumbir a la tentación de una fe ciega en la infalibilidad de estos sistemas. Como toda herramienta poderosa, la inteligencia artificial aplicada a la judicatura encierra riesgos y desafíos que deben ser afrontados con sumo cuidado y diligencia. La opacidad inherente a los algoritmos que sustentan estos modelos, unida a la potencial perpetuación de sesgos históricos o discriminatorios latentes en los datos utilizados para su entrenamiento, nos impelen a erigir robustos mecanismos de control y supervisión que salvaguarden la equidad, imparcialidad y plena sujeción a los principios constitucionales de las decisiones automatizadas.

En este contexto, los recientes avances en el campo de la interpretabilidad de los LLMs, exemplificados por el notable logro del equipo de Anthropic al extraer características monosemánticas del modelo Claude 3 Sonnet, abren una ventana de esperanza para la imprescindible auditoría humana de estos sistemas. Este hito representa un paso crucial hacia la apertura de la caja negra de los LLMs, sentando las bases para el desarrollo de herramientas que

¹⁵² Ibid.

¹⁵³ Juan Gustavo Corvalán, Prometea: Inteligencia Artificial para Transformar Organizaciones Pùblicas (Buenos Aires: Editorial Astrea, 2019), conferencia durante la Asamblea Ordinaria del Consejo Permanente de la Organización de los Estados Americanos, 22 de agosto de 2018, Washington D.C., p. 50. <http://scm.oas.org/pdfs/2018/CP-PRES-CORV.pdf>

permitan monitorizar, corregir y alinear su funcionamiento con los valores y estándares propios de un Estado de Derecho.

Asimismo, la meticulosa selección y continua monitorización de los conjuntos de datos empleados para el entrenamiento de los LLMs emerge como un requisito ineludible para conjurar el riesgo de una justicia sesgada o discriminatoria. Solo mediante la construcción de bases de conocimiento jurisprudencial amplias, diversas y representativas de la pluralidad social podremos asegurar que las decisiones automatizadas reflejen, fielmente, los principios de igualdad y no discriminación consagrados en nuestro ordenamiento jurídico.

En definitiva, la introducción de la inteligencia artificial en la resolución de casos judiciales sencillos y repetitivos constituye una oportunidad extraordinaria para modernizar y optimizar nuestra administración de justicia, pero, también, un desafío formidable que exige una aproximación prudente, garantista y firmemente anclada en los valores constitucionales. Solo así, mediante un delicado equilibrio entre innovación tecnológica y salvaguarda de los derechos fundamentales, podremos materializar el enorme potencial de estos sistemas para coadyuvar a una justicia más ágil, coherente y predecible en los asuntos de menor complejidad, liberando a nuestros jueces y magistrados para un ejercicio más reflexivo y particularizado de su misión en aquellos casos que realmente lo precisen.

- Asistencia/Apoyo en la Administración de Justicia

Junto con la automatización de casos sencillos, la otra gran aplicación de la IA en el ámbito judicial es la de servir de apoyo o asistencia al juez en la resolución de casos complejos. Se trataría de utilizar modelos predictivos o de clasificación para proporcionar al juzgador información relevante sobre el caso, como los factores que estadísticamente se correlacionan con un determinado sentido del fallo, el rango probable de la condena o la indemnización, o los argumentos habitualmente empleados en supuestos análogos. Una información que, lejos de sustituir el criterio jurisdiccional, ha de concebirse como un input más en el proceso deliberativo que culmina con la sentencia.

En este sentido, son ya numerosas las herramientas de "justicia predictiva" que, basadas en técnicas de procesamiento del lenguaje natural y aprendizaje automático, permiten analizar ingentes corpus de resoluciones judiciales para detectar patrones y correlaciones. Así, sistemas

como el estadounidense Lex Machina¹⁵⁴ o el francés Previstico¹⁵⁵ son capaces de predecir el resultado de un litigio con un elevado grado de fiabilidad, a partir de variables como el tipo de caso, la materia, la cuantía, el órgano judicial o incluso la identidad del juez.

Más allá de la mera predicción del fallo, estos sistemas pueden proporcionar al juez una valiosa orientación sobre los elementos fácticos y jurídicos más relevantes para la resolución del caso. Así, por ejemplo, pueden identificar las circunstancias que, en casos similares, han sido consideradas atenuantes o agravantes de la responsabilidad penal, o los criterios jurisprudenciales seguidos para la cuantificación del daño moral. Una información que, debidamente contextualizada y ponderada por el juzgador, puede contribuir a mejorar la coherencia y previsibilidad de las resoluciones, reduciendo la disparidad de criterios entre órganos y la sensación de lotería judicial.

Especial mención merecen los sistemas de apoyo a la decisión judicial que, además de las funcionalidades predictivas, incorporan modelos argumentativos para generar un esbozo de motivación de la sentencia. Naturalmente, este tipo de herramientas no pretende sustituir la insustituible labor de ponderación y argumentación del juez, sino facilitar y agilizar la redacción de la sentencia, asegurando que no se omita ningún elemento relevante y que se respete la estructura lógica del discurso motivador. En última instancia, correspondería al juzgador revisar y modular el texto propuesto por la máquina, adaptándolo a las singularidades del caso e imprimiéndole su particular estilo argumentativo.

Un paso más allá en el uso de la IA como auxilio a la función jurisdiccional lo representan los sistemas integrales de apoyo a la decisión que abarcan todo el ciclo del proceso, desde la admisión a trámite de la demanda hasta la emisión de la sentencia. Estos sistemas combinan diversas técnicas, como el análisis predictivo, la búsqueda semántica, la generación de hipótesis fácticas o la evaluación probabilística de las pruebas, para asistir al juez en las distintas fases del enjuiciamiento.

Mención aparte merecen los sistemas de evaluación del riesgo, como el polémico COMPAS¹⁵⁶ norteamericano, que, a partir del análisis de las circunstancias personales y sociales

¹⁵⁴ Lex Machina, "Predict the Behavior of Courts, Judges, Lawyers and Parties with Legal Analytics," Lex Machina, acceso 22 de mayo de 2024, <https://www.lexmachina.com>

¹⁵⁵ Predictice, "Accédez à toute l'information juridique," Predictice, consultado el 22 de mayo de 2024, <https://www.predictice.com>

¹⁵⁶ Marcela del Pilar Roa Avella y Jesús Eduardo Sanabria-Moyano, "Uso del algoritmo COMPAS en el proceso penal y los riesgos a los derechos humanos," artículo producto del proyecto INV DER 3159 "Inteligencia Artificial:

del encausado, predice la probabilidad de reincidencia futura. Este tipo de herramientas, utilizadas fundamentalmente en el ámbito penal para la adopción de medidas cautelares o la determinación de la pena, han sido objeto de severas críticas por su potencial sesgo discriminatorio, al basarse en factores el nivel económico o el lugar de residencia, que pueden perpetuar y agravar la desigualdad social.

Sin embargo, también han recibido elogios de expertos que precisan que estos sistemas automatizados de evaluación de riesgos poseen la capacidad de reducir, tanto las tasas, como la duración del encarcelamiento para delincuentes de bajo riesgo. Este enfoque conduce a una disminución de los costos presupuestarios y a una mitigación del daño social¹⁵⁷¹⁵⁸.

Cómo precisa la doctrina norteamericana:

“El atractivo de los sistemas automatizados de evaluación de riesgos es que proponen inyectar objetividad en un sistema de justicia que ha sido comprometido, por demasiado tiempo y demasiadas veces, por fallas humanas. Los defensores de los sistemas automatizados de evaluación de riesgos también afirman que hacen que las sentencias sean más transparentes y racionales.

(...)

Los métodos automatizados (de evaluación) son acreditados con dar sustancia a las decisiones y hacerlas más científicas, auditables y, en consecuencia, conferirles la apariencia de legitimidad. El apoyo para la introducción de herramientas algorítmicas de evaluación de riesgos se basa entonces en la premisa de que mejoran el profesionalismo al mejorar la defendibilidad y la responsabilidad de las decisiones, generando uniformidad entre regiones y jurisdicciones, y manteniendo una percepción de validez científica objetiva”¹⁵⁹.

retos y riesgos de los Derechos Humanos en el Sistema Penal”, financiado por la Vicerrectoría de Investigaciones de la Universidad Militar Nueva Granada, convocatoria Proyectos de Investigación Científica vigencia 2020. <https://www.google.co.cr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiVvNXh3qKGAXUFVTABHUU3A1UQFnoECBQQAQ&url=https%3A%2F%2Fdialnet.unirioja.es%2Fdescarga%2Farticulo%2F8438795.pdf&usg=AOvVaw3xLxCMaUz1OcM7jug0Zmab&opi=89978449>

¹⁵⁷ Shaila Dewan, "Judges Replacing Conjecture With Formula for Bail," The New York Times, 16 de junio de 2015. Recuperado de: . <https://www.nytimes.com/2015/06/27/us/turning-the-granting-of-bail-into-a-science.html>

¹⁵⁸ Anne Milgram, "Why Smart Statistics Are the Key to Fighting Crime," TED@BCG San Francisco, filmado en octubre de 2013, TED Video. Recuperado de: https://www.ted.com/talks/anne_milgram_why_smart_statistics_are_the_key_to_fighting_crime

¹⁵⁹ Willem H. Gravett, "Judicial Decision-Making in the Age of Artificial Intelligence," en Multidisciplinary Perspectives on Artificial Intelligence and the Law, editores Henrique Sousa Antunes, Pedro Miguel Freitas, Arlindo

Por ello, su utilización en sede judicial resulta extremadamente delicada, debiendo en todo caso limitarse a una función meramente orientativa y sujetada a un riguroso control humano.

Otro ejemplo similar es el sistema integral de respuesta judicial Xiao Zhi, utilizado en la Corte Suprema Popular China, el cual, además de realizar este primer filtro, brinda apoyo a los jueces mediante el análisis de los escritos presentados en el caso, resumiendo los puntos de controversia conforme son planteados durante el juicio, evaluando las pruebas, calculando las compensaciones y redactando documentos judiciales sobre la marcha¹⁶⁰.

Uno de ellos es STEVIE¹⁶¹, un sistema de razonamiento legal capaz de generar hipótesis fácticas a partir del material probatorio y evaluar su probabilidad a la luz de las pruebas disponibles. La gran ventaja de STEVIE es su capacidad para proporcionar explicaciones comprensibles de su razonamiento, desglosando la contribución de cada elemento probatorio a la probabilidad de cada hipótesis. Esta transparencia es crucial para que el juez pueda valorar críticamente las sugerencias del sistema y retener el control último sobre la fijación de los hechos.

En esta categoría se enmarcan los proyectos implementados por el Poder Judicial de Costa Rica que, al día de hoy, incorporan herramientas potenciadas por sistemas automatizados de inteligencia artificial, demostrando que esta institución ha sido pionera en la región en la adopción de estas innovadoras tecnologías en el ámbito de la justicia. En concreto, estas son las funcionalidades específicas que la administración de justicia costarricense ha implementado gradualmente en varios planes piloto:

- Predicción de la ejecución presupuestaria mediante el análisis del comportamiento histórico, lo cual permitió un ahorro estimado de casi 380,000 dólares anuales¹⁶².

L. Oliveira, Clara Martins Pereira, Elsa Vaz de Sequeira, y Luís Barreto Xavier (Cham, Suiza: Springer Nature Switzerland AG, 2024), 284-285. <https://doi.org/10.1007/978-3-031-41264-6>

¹⁶⁰ Alena Zhabina, "Cómo la IA de China está automatizando el sistema legal," DW, 20 de enero de 2023, <https://p.dw.com/p/4MUY0>

¹⁶¹ Ephraim Nissan, "Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement," AI & Society 32 (2017): 457.. <https://link.springer.com/article/10.1007/s00146-015-0596-5>

¹⁶² Haideer Miranda Bonilla, (2022) Inteligencia Artificial y justicia en Revista de la Facultad de Derecho de México, Tomo LXXII, No. 284. Recuperado de: <https://www.revistas.unam.mx/index.php/rfdm/article/view/83394>, p. 399.

- Desarrollo de un chatbot para responder consultas frecuentes sobre trámites gestión de forma automatizada las 24 horas del día, con un ahorro aproximado de 13,000 dólares mensuales¹⁶³.
- Proyecto piloto de un sistema de "tipificación de documentos" mediante inteligencia artificial en el Juzgado de Cobros de Pérez Zeledón. Este sistema clasifica y organiza los escritos entrantes de forma automática, sin intervención humana, alcanzando un 80 % de precisión. Ha permitido una mayor celeridad procesal y reducción del circulante judicial en este despacho¹⁶⁴.
- Análisis automatizado de las sentencias de la Sala Constitucional para la elaboración de los Informes del Estado de la Justicia. En el tercer y cuarto informe se utilizó IA para clasificar temáticamente las resoluciones, identificar porcentajes de sentencias estimatorias y desestimatorias, tiempos de respuesta, sentencias referentes y patrones jurisprudenciales¹⁶⁵.

Como puede apreciarse, las posibilidades de integración de la IA en las distintas fases del proceso son muy amplias y prometedoras. Sin embargo, su implantación no está exenta de riesgos y desafíos, como bien apunta el texto de referencia. Uno de los más acuciantes es el del "sesgo de automatización"¹⁶⁶, por el cual el juez puede verse tentado a deferir acríticamente a las recomendaciones del sistema, abdicando de su responsabilidad última de juzgar conforme a su propio criterio.

Para conjurar este riesgo, es crucial que la aplicación de estos sistemas vaya acompañada de una sólida formación de los jueces, tanto en los aspectos técnicos de su funcionamiento como en sus limitaciones e implicaciones éticas. Los jueces deben ser plenamente conscientes de que estos sistemas, por sofisticados que sean, no son infalibles ni omniscientes, sino que están sujetos a sesgos y errores derivados de las limitaciones de sus datos de entrenamiento y de los supuestos

¹⁶³ Ibid.

¹⁶⁴ Ibid, p. 400.

¹⁶⁵ Ibid, p. 397.

¹⁶⁶ Databricks, "What is Automation Bias?" Databricks, <https://www.databricks.com/glossary/automation-bias> (accedido el 9 de junio de 2024).

de sus modelos¹⁶⁷. Por ello, deben mantener siempre una actitud crítica y reflexiva, contrastando las sugerencias de la máquina con su propio análisis de los hechos y del derecho aplicable.

Además, esta formación no debe limitarse a los aspectos técnicos, sino que debe abarcar también una sólida fundamentación iusfilosófica que refuerce el compromiso de los jueces con los valores esenciales que informan nuestro Estado de Derecho. Valores como la independencia judicial, la igualdad ante la ley, la presunción de inocencia o la individualización de la justicia, que no pueden ser sacrificados en aras de la eficiencia tecnológica. Solo desde esta doble capacitación, técnica y axiológica, podrán los jueces aprovechar el potencial de la IA como valioso auxiliar, sin abdicar de su rol como garantes últimos de la justicia en el caso concreto.

- Sobre la Difusa Frontera entre “Sistemas de Decisión” y “Sistemas de Apoyo a la Decisión”

La velocidad con la que se han desarrollado las aplicaciones de inteligencia artificial en la administración de justicia ha traído aparejada una controversia fundamental: ¿hasta qué punto un sistema que “apoya” o “asiste” realmente se distingue de aquel que, en la práctica, “decide” por el juez? En el plano teórico, se asume que un sistema de apoyo a la decisión se limita a proporcionar información, sugerir criterios o presentar patrones estadísticos sin desplazar la facultad decisoria humana. Sin embargo, la praxis judicial revela que tal distinción resulta, en muchas ocasiones, difusa o casi ilusoria.

La complejidad se observa, principalmente, cuando el sistema —sin ostentar formalmente la cualidad de “decisor automático”— desarrolla un análisis jurídico tan exhaustivo que el juez se ve *de facto* inclinado a acoger su propuesta. Pensemos, por ejemplo, en una herramienta capaz de rastrear una ingente cantidad de resoluciones análogas, extraer los fundamentos de derecho predominantes y proyectar conclusiones coherentes sobre la sentencia a dictar. Esa formidable “capacidad” de la IA podría, en la práctica, disincentivar al juzgador a discrepar de lo que ya se presenta como una respuesta verosímilmente fundada, ahorrándole tiempo y esfuerzo. Resulta,

¹⁶⁷ Raquel Borges Blázquez, "El sesgo de la máquina en la toma de decisiones en el proceso penal," IUS ET SCIENTIA 6, no. 2 (2020): 54-71, Universidad de Sevilla, España, <https://revistascientificas.us.es/index.php/ies/article/view/14328/12770>

entonces, muy sencillo que el sistema inicialmente concebido como mero “asistente” acabe ejerciendo una influencia decisiva sobre la resolución definitiva.

Este escenario cobra mayor relevancia cuando el sistema pone a disposición del juez un borrador de resolución, que ya contempla los aspectos nucleares del caso: la selección de hechos relevantes, las normas aplicables y la jurisprudencia que sustenta la posición final. La supuesta “asistencia” se traduce, de hecho, en un alto grado de condicionamiento de la decisión, pues la diligente propuesta que formula la IA —con una aparente objetividad y un respaldo estadístico difícil de refutar— desplaza la iniciativa del juzgador hacia el acto de refrendar, corregir mínimamente o, en raras ocasiones, impugnar el planteamiento técnico. El resultado, en muchos casos, difumina la frontera entre un apoyo genuino y la decisión misma.

Basta pensar también en el impacto psicológico que ejerce un sistema calificado de “expertísimo” al procesar miles de resoluciones similares en cuestión de segundos. La misma retórica que precede a la implementación de dichas herramientas (eficiencia garantizada, predictibilidad elevada, fiabilidad extraordinaria) coadyuva a que el juez perciba una suerte de “responsabilidad invertida”: en lugar de cuestionar la propuesta de la IA, debiera justificar por qué se aparta de ella. De ahí que la línea divisoria entre un simple soporte y la asunción real de la decisión pueda convertirse en una mera formalidad.

Las experiencias de “Smart Courts” en China y de diversos proyectos que combinan análisis jurisprudencial, recuperación de materiales probatorios y propuestas de resolución ilustran con nitidez esta tensión. Pese a que, en el papel, se conciben como complementos para aliviar la sobrecarga de los tribunales, el grado de concreción y minuciosidad de sus recomendaciones a menudo se aproxima tanto al fallo final que la intervención humana —lejos de ser un control exhaustivo— tiende a limitarse a un repaso o validación

Tampoco es un tema aislado de determinadas jurisdicciones: cada vez más tribunales y ministerios de justicia de diversos países exploran soluciones que, con la etiqueta de “asistencia”, acaban resolviendo aspectos cruciales del litigio. Incluso en los pilotos de Prometea en América Latina o los programas digitales del HM Courts & Tribunals Service del Reino Unido, la progresiva sofisticación de la IA hace que la diferencia entre “sugerir” y “decidir” dependa más de la conducta del operador humano que de la arquitectura del sistema. En palabras llanas, si el juez opta sistemáticamente por refrendar lo que la IA expone, el resultado es prácticamente indistinguible de un fallo automatizado.

Desde una perspectiva jurídico-académica, la supuesta frontera entre “sistemas de decisión” y “sistemas de apoyo a la decisión” se atenúa sustancialmente cuando el sistema provee todos los elementos sustantivos para la resolución y el juez, por razones de eficiencia o de aparente rigor técnico, acepta sin mayor reparo la salida propuesta. Aunque formalmente se preserve la autoridad del juzgador, los efectos en el proceso y en la motivación de la resolución pueden ser prácticamente equivalentes a los de un dictamen automatizado. **Esta difuminación reclama, por ende, una atención particular de la doctrina y del legislador, ya que, de continuar creciendo la confianza (o dependencia) en estas herramientas, quedará en entredicho la auténtica naturaleza de la función jurisdiccional.**

➤ **Fiscalización de Sentencias**

Esta fiscalización puede proyectarse en dos direcciones fundamentales. Por un lado, los resultados de los modelos predictivos podrían emplearse para constreñir la discrecionalidad judicial, asegurando que casos sustancialmente análogos reciban una respuesta uniforme por parte de los tribunales. En este supuesto, la decisión última seguiría residenciándose en la figura del juez, pero su criterio se vería sometido a un escrutinio *ex post* a la luz de la solución sugerida por el algoritmo. Un ejemplo paradigmático de esta aproximación lo encontramos en el sistema de "avisos de sentencias fuera de lo normal" implementado en ciertos procesos penales en China, que alerta a las autoridades supervisoras cuando una resolución se aparta significativamente de los parámetros de discrecionalidad inferidos de casos pretéritos¹⁶⁸.

Sin embargo, esta modalidad de fiscalización restrictiva de la discrecionalidad suscita serias objeciones desde el prisma de la independencia judicial y de la justicia del caso concreto. Por un lado, existe el riesgo de que estos sistemas se instrumentalicen por parte del poder público para condicionar o predeterminar el sentido de las resoluciones, socavando la autonomía de criterio que debe presidir la función jurisdiccional en un Estado de Derecho. Por otro, la priorización del pasado sobre las circunstancias particulares del caso puede conducir a una aplicación mecánica y descontextualizada del Derecho, sacrificando la equidad en aras de una uniformidad mal entendida.

Ante estos riesgos, parece razonable circunscribir esta fiscalización restrictiva de la discrecionalidad a aquellos casos manifiestamente sencillos y estandarizados en los que, como

¹⁶⁸ Ibidem. De Asís Pulido, Miguel, p. 303.

sociedad, primemos los valores de seguridad jurídica y previsibilidad sobre los de individualización y adaptación de la norma al supuesto concreto. En estos casos, además, cabría plantear si resulta preferible un sistema que confíe directamente la decisión a la máquina mediante la automatización del razonamiento jurídico (como los expuestos en acápite anteriores), o bien uno que preserve la decisión humana, pero la someta a un control *ex post* mediante la fiscalización algorítmica de las eventuales desviaciones anómalas.

Pero la justicia predictiva ofrece también otra vía prometedora para la fiscalización de las sentencias: la detección de posibles sesgos cognitivos o prejuicios discriminatorios en el razonamiento judicial. A través del análisis masivo de resoluciones, estos sistemas podrían identificar patrones decisarios estadísticamente correlacionados con factores extrajurídicos y potencialmente discriminatorios, como el género, la raza o el estatus socioeconómico de las partes. Esta fiscalización no aspiraría ya a constreñir la discrecionalidad judicial, sino a depurarla de influencias espurias que puedan comprometer la imparcialidad y la igualdad en la aplicación de la ley.

Un precedente ilustrativo de esta posibilidad lo encontramos en el célebre caso del sistema COMPAS en Estados Unidos. Aunque no era su propósito original, este programa de evaluación del riesgo de reincidencia puso de manifiesto la existencia de sesgos raciales sistémicos en las decisiones judiciales, al reproducir pautas discriminatorias contra la población afroamericana presentes en los datos históricos utilizados para su entrenamiento.¹⁶⁹ COMPAS no era en sí mismo racista, pero su funcionamiento revelaba los prejuicios implícitos enquistados en el sistema que le servía de base.

Este potencial de los modelos predictivos para sacar a la luz y objetivar sesgos cognitivos y estereotipos discriminatorios resulta sumamente valioso desde la óptica de la imparcialidad y la igualdad ante la ley. No obstante, su implementación práctica debe rodearse de las debidas garantías para evitar efectos contraproducentes. Así, resultaría inadmisible que estos análisis dieran lugar a la elaboración de perfiles individuales de jueces que asociarán, públicamente, determinados magistrados con ciertos prejuicios o predisposiciones. Además, la mera detección estadística de aparentes sesgos por parte de la máquina no puede considerarse concluyente, sino que debe dar lugar a un análisis contextual pormenorizado que descarte posibles factores explicativos legítimos de las correlaciones observadas.

¹⁶⁹ Ibid, p.304.

8. Epílogo del Capítulo I

La irrupción de los grandes modelos de lenguaje (LLMs, por sus siglas en inglés) en el ámbito judicial, y en particular su potencial aplicación como sistemas de apoyo a la decisión basados en el análisis predictivo de sentencias y normas jurídicas, plantea un escenario tan prometedor como desafiante. Aunque su implementación explícita y regulada es todavía incipiente, lo cierto es que la utilidad y accesibilidad de estas herramientas augura su rápida extensión en la práctica judicial, incluso de manera informal o no declarada.

En efecto, la capacidad de los LLMs para procesar y generar lenguaje natural de forma coherente y contextualizada, así como para identificar patrones y correlaciones en grandes volúmenes de texto, los convierte en un recurso de inestimable valor para el juez enfrentado a la tarea de decidir un caso. Ya sea para obtener una predicción del sentido probable del fallo, para identificar los argumentos más frecuentemente utilizados en supuestos análogos, o para generar borradores de resoluciones adaptados a las circunstancias del caso, estos sistemas ofrecen al juzgador una valiosa asistencia que puede contribuir a fundamentar y agilizar su labor.

La utilización de "machotes" o plantillas preestablecidas para la redacción de sentencias es una práctica ampliamente extendida en el ámbito jurisdiccional. Estos modelos estandarizados, que incluyen una estructura básica con secciones como el encabezado, antecedentes, considerandos y parte resolutiva, han demostrado ser una herramienta valiosa para agilizar la labor de los jueces y garantizar la consistencia formal de las resoluciones. Al proveer un esquema predefinido que asegura el cumplimiento de los requisitos legales y procesales, los machotes permiten a los juzgadores centrarse en la incorporación de los detalles específicos de cada caso, ahorrando tiempo y reduciendo errores u omisiones.

Sin embargo, esta práctica también conlleva el riesgo de una excesiva estandarización de la respuesta judicial, que puede ir en detrimento de la adaptación de la sentencia a las particularidades fácticas y jurídicas de cada litigio. La mera inserción de los datos del caso en una plantilla genérica, sin una reflexión profunda sobre las singularidades que lo distinguen, puede derivar en resoluciones formularias y descontextualizadas, que no hagan honor a la exigencia de individualización y equidad inherente a la función de juzgar.

Es precisamente en este punto donde también la inteligencia artificial, y específicamente los Grandes Modelos de Lenguaje (LLMs), pueden suponer un salto cualitativo en la personalización de la atención jurídica. A diferencia de los machotes estáticos, los LLMs tienen la

capacidad de procesar y generar lenguaje natural de forma dinámica y contextualizada, adaptándose a las circunstancias específicas de cada caso. Mediante el análisis de la demanda, las alegaciones de las partes y el material probatorio, estos sistemas pueden identificar los hechos y fundamentos jurídicos relevantes, y generar un primer borrador de sentencia que, además de respetar la estructura formal requerida, incorpore una motivación singularizada y ajustada a las peculiaridades del litigio.

Esta personalización automatizada de las sentencias presenta múltiples ventajas. Por un lado, permite a los jueces partir de una propuesta inicial que ya ha "digerido" las cuestiones clave del caso, lo que reduce significativamente el tiempo y esfuerzo necesarios para la elaboración de la resolución. Por otro, asegura que ningún elemento relevante sea pasado por alto, al tiempo que favorece una mayor coherencia y exhaustividad en la motivación. Además, al basarse en el análisis de un gran corpus de resoluciones previas, los LLMs pueden enriquecer la fundamentación con referencias a los criterios jurisprudenciales más pertinentes, así como con argumentos y consideraciones que quizás escaparían a la atención del juzgador en una redacción manual.

Conscientes de ello, no es aventurado suponer que un número creciente de jueces esté ya recurriendo, de forma más o menos velada, a LLMs en su quehacer jurisdiccional. Ya sea a través de aplicaciones comerciales, de desarrollos internos en los propios tribunales o incluso de simples consultas a modelos de acceso público, la realidad es que estas herramientas se están abriendo paso como un instrumento cotidiano en la "cocina judicial", al margen de su reconocimiento legal expreso.

Este uso informal e incontrolado de los LLMs en la función jurisdiccional, aunque comprensible por las ventajas que ofrece, suscita, sin embargo, serios interrogantes desde el punto de vista de la transparencia, la igualdad y la propia legitimidad democrática de la justicia. ¿Pueden los justiciables conocer y fiscalizar los criterios en base a los cuales se entrena y aplican estos modelos? ¿Están los LLMs reproduciendo y amplificando sesgos discriminatorios presentes en resoluciones previas? ¿Quién controla la calidad, objetividad y actualización de los datos empleados para su entrenamiento? ¿Se está produciendo una excesiva estandarización de la justicia en detrimento de la equidad del caso concreto?

Para afrontar estos riesgos, resulta imprescindible un doble esfuerzo de reflexión colectiva y regulación normativa. Por un lado, es necesario un debate público, plural y multidisciplinar sobre las condiciones y los límites en los que resulta aceptable el uso de LLMs en contextos judiciales.

Un debate que, partiendo del reconocimiento de su utilidad y progresiva extensión, permita identificar aquellos escenarios en los que su empleo puede reportar mayores beneficios -por ejemplo, en tareas de búsqueda y análisis jurisprudencial, generación de primeros borradores o detección de patrones y tendencias-, al tiempo que se preserva el núcleo esencial de la función judicial como garantía última de la justicia del caso concreto.

Pero este debate social debe traducirse, además, en un marco regulatorio específico que, desde el respeto a los principios y las garantías fundamentales del proceso, establezca los requisitos, condiciones y controles aplicables al desarrollo y utilización de LLMs en el ámbito judicial. Una regulación que, entre otros aspectos, debería asegurar:

- a) La transparencia y auditabilidad de los modelos, de modo que su arquitectura, datos de entrenamiento y lógica de funcionamiento sean accesibles y comprensibles para los operadores jurídicos y el público en general.
- b) La objetividad, calidad y representatividad de los datos utilizados para el entrenamiento de los LLMs, evitando sesgos discriminatorios y asegurando su continua actualización.
- c) La posibilidad de revisión humana de las decisiones o recomendaciones generadas por el modelo, preservando siempre la autonomía última del juez para apartarse motivadamente de estas.
- d) La trazabilidad de las interacciones, de manera que quede constancia del grado de influencia que el LLM ha tenido en cada resolución judicial y de los factores considerados por el modelo.
- e) La capacitación tecnológica y ética de los operadores jurídicos, fomentando una cultura de uso responsable y crítico de estas herramientas.

Exclusivamente, a través de esta gobernanza proactiva y garantista de los LLMs en el ámbito judicial, podremos aprovechar su enorme potencial para mejorar la eficiencia, coherencia y predictibilidad de la justicia, al tiempo que minimizamos sus riesgos y preservamos los valores esenciales del Estado de Derecho. Una tarea compleja y urgente que requiere la implicación de todos los actores públicos y privados concernidos, además de que determinará, en buena medida, la calidad de nuestros sistemas judiciales en las próximas décadas.

Aunque el propósito de este trabajo no es proponer cuáles deben ser los usos de la IA en el ámbito judicial, para entrar al análisis jurídico como tal hay que comprender cuál es el potencial de esta tecnología en la administración de justicia. En este sentido, hemos explorado

tres grandes ámbitos de aplicación: la sustitución del juez en casos sencillos y estandarizados, la asistencia y el apoyo en la resolución de casos complejos y la fiscalización de las sentencias para detectar posibles sesgos o desviaciones.

En cuanto a la sustitución en casos fáciles, es probable que esta implementación resulte la más tardía por lo delicado que sería que el juez, como figura humana, deponga sus competencias en una máquina que no está constitucionalmente habilitada al efecto; requeriría una modificación integral de nuestro diseño constitucional. No obstante, es probable que esto suceda más pronto de lo que imaginamos por las tendencias que se pueden observar en el derecho comparado y en el desarrollo exponencial de las capacidades de los LLMS (Large Language Models). Países como China, Reino Unido o Estonia, ya están experimentando con tribunales inteligentes que automatizan la resolución de litigios menores en materias como el comercio electrónico, la propiedad intelectual o el derecho del consumo, con resultados prometedores en términos de agilidad y descarga de trabajo.

Ante esta realidad emergente, resulta imperativo anticiparse y sentar las bases de un marco normativo sólido que garantice que la eventual implementación de estos sistemas se realice con pleno respeto a los principios y garantías esenciales de nuestro Estado de Derecho. Un marco que asegure la transparencia y auditabilidad de los algoritmos, la calidad y representatividad de los datos de entrenamiento, la posibilidad de revisión humana de las decisiones automatizadas y la preservación del núcleo esencial de la función jurisdiccional como garantía última de la justicia del caso concreto. Solo así podremos asegurar que la eficiencia que promete la automatización no se logre a costa de sacrificar la equidad y legitimidad de la respuesta judicial.

La implementación más cercana es posiblemente la segunda: la asistencia y apoyo en la administración de justicia. De hecho – como se conjecturo líneas arriba - es posible que esta tecnología ya esté siendo usada por muchísimos juegadores nacionales para hacer más eficaz su trabajo, sin manifestarlo públicamente. Sin embargo, para que esta simbiosis entre la inteligencia artificial y la humana sea fructífera y legítima, es imprescindible que vaya acompañada de una sólida capacitación de los jueces en los fundamentos técnicos y éticos de estos sistemas. Solo desde una comprensión cabal de sus potencialidades y limitaciones podrán los juegadores aprovechar su potencial sin abdicar de su responsabilidad última de juzgar conforme con su criterio y conciencia. Una formación que debe abarcar también una reflexión profunda sobre los valores y principios que dan sentido a la función judicial y que no pueden ser sacrificados en aras de la eficiencia

tecnológica. De lo contrario, corremos el riesgo de que el "sesgo de automatización" termine por erosionar la independencia y legitimidad de la labor jurisdiccional.

La tercera propuesta, la fiscalización algorítmica de las sentencias, también es una vía a explorar en el futuro. El potencial de los sistemas de IA para detectar patrones decisarios discriminatorios o sesgados abre una oportunidad extraordinaria para avanzar en los ideales de imparcialidad e igualdad en la aplicación de la ley. La experiencia del sistema COMPAS en Estados Unidos, que reveló la existencia de sesgos raciales sistémicos en las decisiones judiciales, es un ejemplo ilustrativo de cómo la tecnología puede ayudarnos a sacar a la luz y objetivar prejuicios y estereotipos enquistados en nuestra praxis jurídica.

Sin embargo, la implementación de estos sistemas de fiscalización debe rodearse de cautela para evitar que se conviertan en una herramienta de control o presión indebida sobre la independencia judicial. La detección de posibles sesgos debe dar lugar a un análisis contextual riguroso a cargo de instancias especializadas e independientes y sus resultados deben manejarse con la debida confidencialidad y respeto a los derechos de los jueces. En ningún caso, sería admisible que estos análisis dieran lugar a la elaboración de perfiles individuales de magistrados que los asociaran públicamente con determinados prejuicios o predisposiciones. La lucha contra la discriminación no puede llevarse a cabo a costa de estigmatizar o coaccionar a quienes tienen la alta responsabilidad de impartir justicia.

En definitiva, la inteligencia artificial aplicada a la administración de justicia encierra un inmenso potencial transformador, pero, también, desafíos y riesgos de gran calado que es preciso afrontar con responsabilidad y altura de miras. **El reto, en última instancia, consiste en diseñar un modelo de justicia aumentada en el que la tecnología se ponga al servicio de los valores superiores que dan sentido a esta función esencial del Estado de Derecho. Un modelo en el que la eficiencia se conjugue con la equidad, la coherencia con la adaptación al caso concreto, la predictibilidad con la capacidad de innovación jurídica.**

Para ello, será imprescindible un diálogo interdisciplinar y una colaboración estrecha entre todos los actores implicados: judicatura, abogacía, academia, expertos tecnológicos, responsables públicos y sociedad civil. Solo desde una reflexión compartida y un compromiso colectivo podremos alumbrar un nuevo paradigma de justicia que, sin renunciar a sus principios fundamentales, sepa aprovechar las inmensas posibilidades de la revolución digital para hacer realidad una tutela judicial más accesible, ágil y fiable para todos los ciudadanos.

El horizonte que se vislumbra es tan prometedor como desafiante. Pero si algo nos enseña la historia del Derecho es que sus categorías e instituciones nunca han sido inmutables, sino que han evolucionado al compás de las transformaciones sociales, económicas y tecnológicas de cada época. La irrupción de la inteligencia artificial en el mundo del Derecho no es sino un nuevo capítulo de esta evolución, que nos interpela a repensar y actualizar nuestra concepción de la Justicia para adaptarla a los retos y oportunidades del siglo XXI.

La justicia del futuro será, sin duda, una justicia tecnológicamente aumentada. Pero solo será una justicia digna de tal nombre si sabe mantener en su centro la dignidad inviolable de cada persona, la igualdad radical de todos ante la ley y la sensibilidad para las circunstancias únicas e irrepetibles de cada caso. Una justicia que, en el maridaje entre la máquina y el hombre, sepa preservar siempre la primacía de la conciencia sobre el algoritmo.

En este contexto, resulta imprescindible dirigir nuestra mirada hacia las experiencias y los avances normativos que se están produciendo en otros ordenamientos jurídicos, especialmente en el ámbito de la Unión Europea. La regulación de la inteligencia artificial aplicada a la administración de justicia es un reto de alcance global, que trasciende las fronteras de los Estados y exige una respuesta coordinada y armonizada a nivel supranacional.

La Unión Europea, consciente de la trascendencia de este desafío, está liderando los esfuerzos por construir un marco normativo que permita aprovechar el potencial de la IA en la justicia al tiempo que se salvaguardan los derechos fundamentales y los principios del Estado de Derecho. Por ello, en los siguientes capítulos abordaremos un análisis pormenorizado de la normativa relevante en la Unión Europea en materia de IA y justicia. Nuestro objetivo será doble: por un lado, diagnosticar la situación de nuestro propio ordenamiento jurídico a la luz de estos desarrollos, identificando posibles lagunas, inconsistencias o necesidades de adaptación; por otro, extraer las mejores prácticas y orientaciones que puedan servir de inspiración para una regulación nacional a la altura de los retos que plantea esta nueva frontera tecnológica.

Únicamente desde el estudio riguroso de las experiencias comparadas y la reflexión crítica sobre nuestro propio marco normativo, podremos avanzar hacia ese horizonte de una justicia aumentada que sepa conjugar lo mejor de la inteligencia humana y artificial al servicio de los valores superiores de nuestro Estado de Derecho. Un horizonte que ya se vislumbra en el trabajo pionero de las instituciones europeas y que nos corresponde a todos, como juristas y como ciudadanos, hacer realidad también en nuestro país.

Capítulo II. La Regulación de la Inteligencia Artificial en la Administración de Justicia: un Análisis Exhaustivo del Marco Normativo Europeo

2.1.- La Necesidad de un Marco Regulatorio para la IA en la Justicia

En el alba de la cuarta revolución industrial, nos encontramos ante un paradigma tecnológico que promete transformar los cimientos mismos de nuestras instituciones jurídicas. La inteligencia artificial (IA), otrora confinada al ámbito de la ciencia ficción, se erige hoy como una realidad tangible y omnipresente, cuya incursión en los recintos de la administración de justicia suscita tanto fascinación como inquietud.

La implementación de sistemas automatizados de decisión basados en IA en el ámbito judicial no es una mera especulación futurista, sino un fenómeno *in fieri* que demanda, con urgencia ineluctable, la articulación de un marco regulatorio robusto y coherente. La ausencia de tal andamiaje normativo podría desembocar en una peligrosa anomía digital, donde los algoritmos, cual modernos oráculos, dictaminen sobre derechos y libertades fundamentales sin el debido escrutinio jurídico y ético.

La necesidad imperiosa de un marco regulatorio específico para la IA en la justicia se fundamenta en las siguientes consideraciones fundamentales:

2.1.1.- Salvaguardia de los Principios Fundamentales del Estado de Derecho

La introducción de sistemas de IA en los procesos decisorios judiciales amenaza con subvertir principios cardinales como la igualdad ante la ley, la seguridad jurídica y el derecho a un juicio justo. Esta amenaza no es meramente hipotética, sino que se materializa en riesgos concretos y cuantificables.

En primer lugar, la igualdad ante la ley, consagrada en el artículo 20 de la Carta de los Derechos Fundamentales de la Unión Europea¹⁷⁰, podría verse comprometida por los sesgos inherentes a los algoritmos de aprendizaje automático. Estos sesgos, que pueden ser de naturaleza estadística, social o incluso técnica, tienen el potencial de perpetuar y amplificar discriminaciones preexistentes en el sistema judicial¹⁷¹. Cómo se esbozó brevemente en acápite anteriores, un estudio realizado por ProPublica en 2016 sobre el sistema COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), utilizado en varios estados de EE.UU. para evaluar el riesgo de reincidencia, reveló un sesgo racial significativo en sus predicciones, con una tasa de falsos positivos dos veces mayor para los acusados afroamericanos que para los blancos¹⁷².

En segundo lugar, la seguridad jurídica, principio fundamental del ordenamiento jurídico europeo reconocido por el Tribunal de Justicia de la Unión Europea¹⁷³, se ve amenazada por la opacidad e imprevisibilidad de ciertos sistemas de IA. La complejidad de los algoritmos de aprendizaje profundo, en particular, puede resultar en decisiones cuya lógica subyacente es ininteligible incluso para sus propios diseñadores, lo cual ha llevado a algunos autores a hablar de una "caja negra algorítmica". Esta opacidad no solo dificulta el escrutinio judicial de las decisiones basadas en IA, sino que, también, socava la previsibilidad del derecho, elemento esencial de la seguridad jurídica.

Finalmente, el derecho a un juicio justo, consagrado en el artículo 6 del Convenio Europeo de Derechos Humanos¹⁷⁴, se ve potencialmente comprometido por la introducción de sistemas de IA en el proceso judicial. La utilización de algoritmos predictivos en la toma de decisiones judiciales, como la determinación de la prisión preventiva o la cuantificación de penas, plantea serias dudas sobre la compatibilidad de estas prácticas con las garantías procesales fundamentales. En particular, el derecho a ser juzgado por un tribunal independiente e imparcial podría verse

¹⁷⁰ Unión Europea. "Carta de los Derechos Fundamentales de la Unión Europea." Diario Oficial de la Unión Europea, C 326, 26 de octubre de 2012. <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:12012P/TXT&from=EN>

¹⁷¹ Barocas, Solon, y Andrew D. Selbst. "Big Data's Disparate Impact." California Law Review 104, no. 3 (2016): 671-732. <https://www.californialawreview.org/print/2-big-data-disparate-impact/>, p. 674.

¹⁷² Angwin, Julia, et al. "Machine Bias." ProPublica, 23 de mayo de 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

¹⁷³ Tribunal de Justicia de la Unión Europea. Sentencia de 21 de septiembre de 1983, Deutsche Milchkontor GmbH y otros contra República Federal de Alemania, asuntos acumulados 205 a 215/82, ECLI:EU:C:1983:233,

¹⁷⁴ Consejo de Europa. "Convenio Europeo de Derechos Humanos." 1950. https://www.echr.coe.int/documents/convention_spa.pdf

menoscabado si los jueces humanos se ven indebidamente influenciados por las recomendaciones de los sistemas de IA, fenómeno conocido como "sesgo de automatización"⁷.

2.1.2.- Transparencia y Rendición de Cuentas

La opacidad inherente a ciertos algoritmos de aprendizaje automático plantea desafíos significativos en términos de transparencia y rendición de cuentas, principios fundamentales del Estado de Derecho y requisitos esenciales para la legitimidad del sistema judicial.

El derecho a una tutela judicial efectiva, reconocido en el artículo 47 de la Carta de los Derechos Fundamentales de la UE¹⁷⁵, exige que las decisiones judiciales sean motivadas y susceptibles de revisión. Sin embargo, la complejidad de los modelos de IA avanzados puede resultar en decisiones cuyo proceso de razonamiento es inaccesible no solo para los justiciables, sino, incluso, para los propios jueces que las adoptan.

Este fenómeno, denominado por algunos autores como "cajas negras algorítmicas", plantea serios interrogantes sobre la posibilidad de un control judicial efectivo de las decisiones basadas en IA. ¿Cómo puede un tribunal de apelación revisar una decisión cuya lógica subyacente es ininteligible? ¿Cómo puede un ciudadano impugnar una sentencia basada en un algoritmo cuyo funcionamiento desconoce?

La falta de transparencia algorítmica no solo compromete el derecho a la tutela judicial efectiva, sino que, también, socava la confianza pública en el sistema judicial, elemento esencial para la legitimidad del Estado de Derecho. Un estudio realizado por el Consejo de Europa en 2018 reveló que la mayoría de los ciudadanos europeos se muestran escépticos ante la idea de que las

¹⁷⁵ **Artículo 47. Derecho a la tutela judicial efectiva y a un juez imparcial.** Toda persona cuyos derechos y libertades garantizados por el Derecho de la Unión hayan sido violados tiene derecho a la tutela judicial efectiva respetando las condiciones establecidas en el presente artículo. Toda persona tiene derecho a que su causa sea oída equitativa y públicamente y dentro de un plazo razonable por un juez independiente e imparcial, establecido previamente por la ley. Toda persona podrá hacerse aconsejar, defender y representar. Se prestará asistencia jurídica gratuita a quienes no dispongan de recursos suficientes siempre y cuando dicha asistencia sea necesaria para garantizar la efectividad del acceso a la justicia.

decisiones judiciales sean tomadas por algoritmos, precisamente debido a preocupaciones sobre la falta de transparencia y rendición de cuentas¹⁷⁶.

2.1.3.- Consideraciones Éticas

La implementación de sistemas de IA en la administración de justicia suscita interrogantes éticos de profundo calado, que trascienden el ámbito puramente jurídico para adentrarse en cuestiones fundamentales de filosofía moral y política.

En primer lugar, la utilización de algoritmos predictivos en la toma de decisiones judiciales plantea la cuestión del determinismo tecnológico. Por ejemplo, ¿en qué medida la predicción algorítmica de la probabilidad de reincidencia de un acusado puede convertirse en una profecía autocumplida? El filósofo del derecho Mireille Hildebrandt advierte sobre el riesgo de que los sistemas de IA, al predecir el comportamiento futuro basándose en datos históricos, puedan perpetuar y amplificar las desigualdades e injusticias existentes en el sistema judicial¹⁷⁷.

En segundo lugar, la implementación de sistemas de IA en la justicia plantea cuestiones fundamentales sobre la naturaleza de la justicia y el papel del juicio humano en su administración. ¿Puede un algoritmo captar las sutilezas y complejidades del comportamiento humano que un juez experimentado tomaría en cuenta? ¿Cómo se ponderan valores como la equidad y la misericordia en un sistema algorítmico?

Finalmente, la utilización de IA en la justicia suscita preocupaciones sobre la dignidad humana y la autonomía individual por cuanto la reducción de los individuos a conjuntos de datos procesables algorítmicamente podría constituir una forma de "cosificación" que atenta contra la dignidad humana. En el contexto judicial, esto plantea la cuestión de si es éticamente aceptable que decisiones que afectan profundamente a la vida y libertad de los individuos sean tomadas, aunque sea parcialmente, por máquinas.

¹⁷⁶ Consejo de Europa. "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment." Adoptada en la 31^a reunión plenaria de la CEPEJ, Estrasburgo, 3-4 de diciembre de 2018. <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>, p. 14.

¹⁷⁷ Hildebrandt, Mireille. "Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics." University of Toronto Law Journal 68, no. supplement 1 (2018): 12-35. <https://www.utpjournals.press/doi/full/10.3138/utlj.2017-0044>, p. 33.

2.2.- El Papel Pionero de la Unión Europea en la Gobernanza de la IA

En esta nueva era definida por la omnipresencia de la inteligencia artificial, la Unión Europea ha asumido, con visión prospectiva el papel de arquitecta pionera de un marco regulatorio comprehensivo para esta tecnología disruptiva. El Reglamento sobre Inteligencia Artificial (en adelante, AI Act), propuesto por la Comisión Europea en abril de 2021¹, se erige como la piedra angular de este ambicioso proyecto normativo, sin parangón en el panorama jurídico internacional.

Este corpus iuris no surge *ex nihilo*, sino que se inscribe en una tradición jurídica que ha situado históricamente a la UE a la vanguardia de la protección de los derechos fundamentales en la era digital. El Reglamento General de Protección de Datos (RGPD)¹⁷⁸, paradigma de esta tradición, ha demostrado la capacidad de la Unión para establecer estándares globales en materia de regulación tecnológica, proyectando su influencia mucho más allá de las fronteras comunitarias.

2.3. The EU AI Act

El Reglamento Europeo sobre Inteligencia Artificial (IA) representa la culminación de un proceso regulatorio sin precedentes en el ámbito global. Para comprender su génesis y significado, es imperativo contextualizar su desarrollo en el marco más amplio de los enfoques regulatorios internacionales y la evolución del posicionamiento de la Unión Europea frente a los desafíos y las oportunidades que presenta la IA.

A nivel de Derecho comparado, la regulación jurídica de la IA puede clasificarse en tres modelos paradigmáticos. Un primer grupo de Estados, ejemplificado por Estados Unidos y Reino Unido, ha optado por un enfoque basado en el liberalismo del propio mercado. Este modelo se caracteriza por una intervención regulatoria mínima, confiando en la capacidad de autorregulación de la industria y en los mecanismos de mercado para abordar los potenciales riesgos asociados a la IA. En el extremo opuesto, un segundo grupo de Estados, con China y Rusia como exponentes principales, ha priorizado un enfoque imperativo basado en el papel director del Estado, donde la

¹⁷⁸ 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, <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32016R0679&from=ES>

regulación de la IA se enmarca en una estrategia más amplia de control y direccionamiento estatal de la innovación tecnológica¹⁷⁹.

En este contexto, la Unión Europea ha emergido como abanderada de un tercer paradigma regulatorio, caracterizado por la protección de los derechos fundamentales como piedra angular de su aproximación a la IA. Este enfoque, que podríamos denominar "antropocéntrico", busca equilibrar el fomento de la innovación tecnológica con la salvaguarda de los valores y principios fundamentales sobre los que se erige el proyecto europeo.

La génesis de este enfoque puede rastrearse hasta el Libro Blanco sobre Inteligencia Artificial, publicado por la Comisión Europea el 19 de febrero de 2020. Este documento seminal esbozó los contornos de lo que se convertiría en el paradigma regulatorio europeo en materia de IA, articulando una visión de "excelencia y confianza" que buscaba posicionar a Europa como líder global en innovación tecnológica responsable. El Libro Blanco no solo abordó los riesgos asociados a determinados usos de esta tecnología disruptiva, sino que, también, propuso un enfoque regulatorio basado en el riesgo, sentando las bases conceptuales para el futuro reglamento.

Previo a la publicación del Libro Blanco, la Comisión Europea ya había dado pasos significativos hacia la articulación de un marco ético para la IA. El 8 de abril de 2019, se publicaron las "Directrices éticas para una IA fiable"¹⁸⁰, elaboradas por el Grupo de Expertos de Alto Nivel sobre IA de la Comisión Europea. Estas directrices, aunque no vinculantes, establecieron requisitos clave que los sistemas de IA deberían cumplir para ser considerados fiables, entre los que se incluían la transparencia, la diversidad y la no discriminación y la supervisión humana. Seguidamente, el 26 de junio de 2019 el mismo grupo emitió el documento denominado "Recomendaciones de Política e Inversión para una IA Confiable"¹⁸¹, cuyo enfoque gravita en cuatro áreas principales: empoderar y proteger a los humanos y la sociedad, transformar el sector

¹⁷⁹ Moisés Barrio Andrés, dir., *El Reglamento Europeo de Inteligencia Artificial* (Valencia: Tirant lo Blanch, 2024), p. 27, ISBN 978-84-1071-304-8.

¹⁸⁰ Comisión Europea, *Directrices éticas para una IA fiable*, elaborado por el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial, publicado el 8 de abril de 2019. Recuperadas de: <https://digital-strategy.ec.europa.eu/es/library/ethics-guidelines-trustworthy-ai>

¹⁸¹ Comisión Europea, *Recomendaciones de política e inversión para una IA fiable*, elaborado por el Grupo de Expertos de Alto Nivel en Inteligencia Artificial, publicado el 26 de junio de 2019. Recuperado de: <https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence>

privado, el sector público como catalizador de crecimiento sostenible e innovación y, así, asegurar capacidades de investigación de clase mundial.

La publicación de estas directrices marcó un hito crucial en la aproximación europea a la IA al establecer un marco ético que, posteriormente, guiaría el desarrollo de una legislación vinculante. Tras este enfoque materializado a través del soft law, la Comisión Europea finalmente adoptó una postura jurídica, solicitando la adopción de normas legales armonizadas para el desarrollo, comercialización y uso de sistemas de IA.

De manera complementaria, el Parlamento Europeo, en ejercicio de sus facultades consultivas y de iniciativa legislativa, instó a la Comisión Europea a efectuar una evaluación exhaustiva del impacto de la Inteligencia Artificial (IA) y a formular un marco regulatorio comprehensivo a nivel de la Unión Europea (UE) en materia de IA. Estas directrices se plasmaron en las extensas recomendaciones emitidas el 16 de febrero de 2017, concernientes a las normas de Derecho Civil sobre robótica (2015/2103(INL))¹⁸².

Subsecuentemente, en el transcurso del 2020, el órgano legislativo de la Unión adoptó un conjunto de resoluciones que abordaban aspectos cruciales relacionados con la IA, a saber: consideraciones éticas (2020/2012(INL))¹⁸³, régimen de responsabilidad civil (2020/2014(INL))¹⁸⁴ y protección de los derechos de propiedad intelectual (2020/2015(INI))¹⁸⁵. Este corpus legislativo fue complementado en 2021 con la promulgación de resoluciones adicionales que versaban sobre la implementación de la IA en diversos ámbitos, incluyendo las esferas civil y militar (2020/2013(INI))¹⁸⁶, los sectores educativo, cultural y audiovisual

¹⁸² Parlamento Europeo, *Resolución de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103(INL))*. Recuperado de: https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_ES.html

¹⁸³ Parlamento Europeo, *Resolución de 20 de octubre de 2020, con recomendaciones destinadas a la Comisión sobre un marco de los aspectos éticos de la inteligencia artificial, la robótica y las tecnologías conexas (2020/2012(INL))*. Recuperado de: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_ES.html

¹⁸⁴ Parlamento Europeo, *Resolución de 20 de octubre de 2020, con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial (2020/2014(INL))*. Recuperado de: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_ES.html

¹⁸⁵ Parlamento Europeo, *Resolución de 20 de octubre de 2020, sobre los derechos de propiedad intelectual para el desarrollo de las tecnologías relativas a la inteligencia artificial (2020/2015(INI))*. Recuperado de: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_ES.html

¹⁸⁶ Parlamento Europeo, *Resolución de 20 de enero de 2021, sobre inteligencia artificial: cuestiones de interpretación y de aplicación del Derecho internacional en la medida en que la UE se ve afectada en los ámbitos de los usos civil*

(2020/2017(INI))¹⁸⁷, así como su aplicación en el contexto del Derecho Penal y su uso por parte de las autoridades policiales y judiciales en asuntos penales (2020/2016(INI))¹⁸⁸.

La Comisión Europea, respondiendo entonces a estos llamados, puso en marcha una amplia consulta pública en 2020, que culminó el 21 de abril de 2021 con la publicación de una evaluación de impacto¹⁸⁹, un estudio apoyando dicha evaluación¹⁹⁰ y la propuesta legislativa para el Reglamento sobre IA. Esta propuesta, que recibió las observaciones de las partes interesadas, identificó varios problemas planteados por el desarrollo y uso de los sistemas de IA, atendiendo a sus características específicas: la opacidad, complejidad, imprevisibilidad y el comportamiento parcialmente autónomo.

El proceso legislativo del Reglamento de IA ha sido particularmente intenso y complejo, reflejando la importancia y el carácter controvertido de la materia¹⁹¹. Sobre la base de la propuesta de la Comisión, el Consejo adoptó su posición común el 6 de diciembre de 2022. En el Parlamento, el expediente se asignó conjuntamente a la Comisión de Mercado Interior y Protección del Consumidor (IMCO) y a la Comisión de Libertades Civiles, Justicia y Asuntos de Interior (LIBE).

El Parlamento adoptó su posición negociadora el 14 de junio de 2023 (con 499 votos a favor, 28 en contra y 93 abstenciones), introduciendo enmiendas sustanciales al texto de la Comisión. Tras largas negociaciones entre el Consejo y el Parlamento Europeo, se alcanzó un acuerdo provisional sobre el RIA el 9 de diciembre de 2023. Las comisiones LIBE e IMCO del

y militar, así como de la autoridad del Estado fuera del ámbito de la justicia penal (2020/2013(INI)). Recuperado de: https://www.europarl.europa.eu/doceo/document/TA-9-2021-0009_ES.html

¹⁸⁷ Parlamento Europeo, *Resolución de 19 de mayo de 2021, sobre la inteligencia artificial en los sectores educativo, cultural y audiovisual* (2020/2017(INI)). Recuperado de: https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_ES.html

¹⁸⁸ Parlamento Europeo, Resolución de 6 de octubre de 2021, sobre la inteligencia artificial en el Derecho penal y su utilización por las autoridades policiales y judiciales en asuntos penales (2020/2016(INI)). Recuperado de: https://www.europarl.europa.eu/doceo/document/TA-9-2021-0405_ES.html

¹⁸⁹ Comisión Europea, *Evaluación de impacto: Acompañando la Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas sobre inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión*, SWD(2021) 84 final (Bruselas, 21 de abril de 2021). Recuperado de: <https://digital-strategy.ec.europa.eu/en/library/impact-assessment-regulation-artificial-intelligence>

¹⁹⁰ Comisión Europea, *Estudio para apoyar una evaluación de impacto de los requisitos normativos para la Inteligencia Artificial en Europa. Informe final* (Publicaciones de la UE), completado para la Comisión Europea, DG CONNECT, Bruselas, Abril del 2021. Recuperado de: <https://op.europa.eu/en/publication-detail/-/publication/55538b70-a638-11eb-9585-01aa75ed71a1/language-en>

¹⁹¹ Barrio Andrés, *El Reglamento Europeo de Inteligencia Artificial*, p. 28-30.

Parlamento Europeo aprobaron el texto final en una votación conjunta el 13 de febrero de 2024, con una mayoría abrumadora (71 votos a favor, 8 votos en contra y 7 abstenciones).

El Parlamento Europeo aprobó el texto en su sesión plenaria el 13 de marzo de 2024 (con 523 votos a favor, 46 votos en contra y 46 abstenciones), seguida de una corrección de errores finalmente aprobada con fecha 25 de abril de 2024. El Consejo adoptó formalmente el texto con fecha 21 de mayo de 2024 y fue publicado en el Diario Oficial de la Unión Europea (DOUE) con fecha 12 de julio de 2024.

La aplicación del Reglamento seguirá un calendario escalonado, reflejando la complejidad de la materia y la necesidad de proporcionar a los actores implicados tiempo suficiente para adaptarse a las nuevas exigencias normativas. Como regla general, será aplicable a los 24 meses de su entrada en vigor, aunque ciertos capítulos y disposiciones tendrán plazos de aplicación diferentes, que van desde los 6 hasta los 36 meses.

En este contexto de liderazgo normativo de la Unión Europea en el ámbito digital, el AI Act se presenta como la cristalización más avanzada de la visión regulatoria comunitaria en materia de inteligencia artificial. Trasciende la mera continuidad con iniciativas previas como el RGPD, para erigirse en un nuevo paradigma regulatorio que aspira a establecer un equilibrio entre la innovación tecnológica y la salvaguarda de los derechos fundamentales. La materialización de esta ambición se plasma en la naturaleza jurídica específica del AI Act, cuya forma y fundamento legal reflejan la complejidad y el alcance de los desafíos que pretende abordar. Analicemos, pues, las características jurídicas distintivas de este instrumento normativo que lo convierten en un hito en la gobernanza global de la IA.

2.3.1.- Naturaleza Jurídica y Ámbito de Aplicación

En primer término, es menester subrayar que el AI Act adopta la forma de Reglamento, instrumento jurídico contemplado en el artículo 288 del Tratado de Funcionamiento de la Unión Europea (TFUE):

“ACTOS JURÍDICOS DE LA UNIÓN. Artículo 288 (antiguo artículo 249 TCE) Para ejercer las competencias de la Unión, las instituciones adoptarán reglamentos, directivas, decisiones, recomendaciones y dictámenes. El reglamento tendrá un alcance general.

*Será obligatorio en todos sus elementos y directamente aplicable en cada Estado miembro. La directiva obligará al Estado miembro destinatario en cuanto al resultado que deba conseguirse, dejando, sin embargo, a las autoridades nacionales la elección de la forma y de los medios. La decisión será obligatoria en todos sus elementos. Cuando designe destinatarios, sólo será obligatoria para éstos. Las recomendaciones y los dictámenes no serán vinculantes*¹⁹². (Énfasis agregado)

Esta elección no es baladí, pues implica su alcance general, obligatoriedad en todos sus elementos y aplicabilidad directa en cada Estado miembro. A diferencia de las directivas, que requieren transposición al ordenamiento jurídico nacional, el reglamento se integra automáticamente en el acervo normativo de los Estados miembros desde su entrada en vigor.

La base jurídica sobre la cual se erige el AI Act es dual, fundamentándose en los artículos 16 y 114 del TFUE. La base jurídica sobre la cual se erige el AI Act es dual, fundamentándose en los artículos 16 y 114 del TFUE. El artículo 16 confiere a la Unión la competencia para establecer normas relativas a la protección de las personas físicas respecto del tratamiento de datos personales, mientras que el artículo 114 faculta al Parlamento Europeo y al Consejo para adoptar medidas relativas a la aproximación de las legislaciones nacionales que tengan por objeto el establecimiento y el funcionamiento del mercado interior. Esta doble fundamentación jurídica refleja la naturaleza poliédrica de la regulación sobre IA, que abarca, tanto aspectos de protección de derechos fundamentales, como de armonización del mercado único digital.

El ámbito de aplicación *ratione materiae* del AI Act es notablemente amplio, abarcando el desarrollo, comercialización y uso de sistemas de IA en la Unión, independientemente de si los proveedores están establecidos dentro o fuera de la UE. Esta extraterritorialidad es un rasgo distintivo que subraya la vocación del Reglamento de erigirse como un estándar global, fenómeno conocido en la doctrina como el "efecto Bruselas", que se examinara más adelante.

En cuanto a su relación con otras normativas de la UE, el AI Act se concibe como *lex specialis* respecto a las normas horizontales del Derecho de la Unión en materia de protección de datos personales, protección de los consumidores y seguridad de los productos. No obstante, el

¹⁹² Versión Consolidada del Tratado de Funcionamiento de la Unión Europea, art. 288, 2012 O.J. C 326/47, <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A12012E%2FTXT>

reglamento establece cláusulas de articulación para garantizar su coherencia con el acervo comunitario, particularmente con el Reglamento General de Protección de Datos (RGPD).

La naturaleza jurídica del AI Act, como reglamento, implica importantes consecuencias para los Estados miembros. Por un lado, limita su margen de maniobra en la implementación, al no requerir ni permitir medidas nacionales de transposición. Por otro, impone obligaciones directas a las autoridades nacionales, que deberán adaptar sus estructuras administrativas y judiciales para garantizar la aplicación efectiva del reglamento.

Así las cosas, es claro que la naturaleza jurídica del AI Act como reglamento de la UE, fundamentado en una base legal dual y con un amplio ámbito de aplicación, lo convierte en un instrumento jurídico de singular potencia normativa. Su carácter directamente aplicable y su vocación de estándar global en la regulación de la IA auguran un impacto profundo no solo en el ordenamiento jurídico de la UE, sino en el panorama regulatorio internacional de las tecnologías emergentes.

Como se esbozaba supra, el AI Act, en su concepción y alcance, trasciende la mera respuesta reactiva a los desafíos planteados por la IA, para erigirse en un verdadero proyecto de ingeniería jurídica que aspira a moldear el desarrollo futuro de esta tecnología. En palabras del legislador europeo, el objetivo es "*mejorar el funcionamiento del mercado interior mediante el establecimiento de un marco jurídico uniforme para el desarrollo, la comercialización y el uso de la inteligencia artificial de conformidad con los valores de la Unión*"¹⁹³.

La relevancia de este instrumento normativo en el panorama regulatorio global es difícil de sobreestimar. En un contexto donde la IA se perfila como la tecnología definitoria del siglo XXI, con implicaciones profundas en todos los ámbitos de la vida social y económica, el AI Act se presenta como el primer intento sistemático y holístico de establecer reglas del juego claras y vinculantes para su desarrollo y despliegue.

¹⁹³ Comisión Europea. "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." COM(2021) 206 final, 21 de abril de 2021, p. 20. <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52021PC0206&from=EN>

En el contexto específico de la administración de justicia, ámbito neurálgico del Estado de Derecho y piedra de toque de toda sociedad democrática, el AI Act reviste una relevancia capital. La implementación de sistemas automatizados de decisión basados en IA en el ámbito judicial plantea desafíos jurídicos, éticos y filosóficos de una profundidad y complejidad sin precedentes, que el legislador europeo ha abordado con un grado de minuciosidad y rigor técnico encomiables.

El subsiguiente análisis se propone, pues, examinar pormenorizadamente las disposiciones del AI Act relativas a la implementación de sistemas de IA en la administración de justicia, a la luz de los principios fundamentales que informan el enfoque regulatorio europeo. Este ejercicio no solo reviste un interés académico, sino que se perfila como una herramienta invaluable para comprender las implicaciones jurídicas y prácticas de la introducción de tecnologías de IA en el sistema judicial.

2.3.2.- Principios Fundamentales del Enfoque Regulatorio Europeo en Materia de Inteligencia Artificial

El enfoque regulatorio europeo en materia de inteligencia artificial, cristalizado en el AI Act, se articula en torno a una constelación de principios fundamentales que, en su conjunción, configuran un paradigma normativo sin precedentes en el panorama jurídico internacional. Estos principios, que a continuación se analizarán con la profundidad que su trascendencia merece, no son meras declaraciones programáticas, sino que impregnan la totalidad del texto normativo, dotándolo de una coherencia interna y una visión estratégica que lo distinguen de otros intentos regulatorios en la materia.

a. Antropocentrismo Regulatorio

El principio de antropocentrismo regulatorio se erige como la piedra angular sobre la que se edifica todo el edificio normativo del AI Act. Este principio postula que el desarrollo, despliegue y utilización de sistemas de inteligencia artificial debe estar inexorablemente supeditado al bienestar humano y al respeto irrestricto de los derechos fundamentales.

Esta concepción antropocéntrica encuentra su expresión más diáfana en el artículo 1 del reglamento, que establece como objetivo primordial "*garantizar un alto nivel de protección de la*

salud, la seguridad y los derechos fundamentales de las personas"¹. No se trata, pues, de una mera declaración de intenciones, sino de un mandato imperativo que permea todo el texto normativo y que se traduce en disposiciones concretas y vinculantes.

Así, el artículo 5 del AI Act prohíbe taxativamente una serie de prácticas de IA consideradas incompatibles con los valores fundamentales de la Unión. Entre estas prohibiciones destacan:

1. **La utilización de sistemas de IA que empleen técnicas subliminales para distorsionar materialmente el comportamiento de una persona de manera que pueda causarle perjuicios físicos o psicológicos:** (a) la comercialización, puesta en servicio o uso de un sistema de IA que emplee técnicas subliminales más allá de la conciencia de una persona o técnicas intencionalmente manipulativas o engañosas, con el objetivo o el efecto de distorsionar materialmente el comportamiento de una persona, o un grupo de personas, al afectar apreciablemente su capacidad para tomar una decisión informada, causando así que tomen una decisión que de otra manera no habrían tomado de manera que cause o sea razonablemente probable que cause daño significativo a esa persona, otra persona o grupo de personas.
2. **La explotación de vulnerabilidades específicas de determinados grupos de personas debido a su edad, discapacidad física o mental:** (b) la comercialización, puesta en servicio o uso de un sistema de IA que explote cualquiera de las vulnerabilidades de una persona natural o un grupo específico de personas debido a su edad, discapacidad o una situación social o económica específica, con el objetivo o el efecto de distorsionar materialmente el comportamiento de esa persona o de una persona perteneciente a ese grupo de manera que cause o sea razonablemente probable que cause daño significativo a esa persona o a otra persona.
3. **El uso de sistemas de puntuación social por parte de las autoridades públicas:** (c) la comercialización, puesta en servicio o uso de sistemas de IA para la evaluación o clasificación de personas naturales o grupos de personas durante un cierto período de tiempo basado en su comportamiento social o características personales conocidas, inferidas o predichas, con el puntaje social que conduzca a cualquiera o ambas de las siguientes situaciones: (i) Tratamiento perjudicial o desfavorable de ciertas personas

naturales o grupos de personas en contextos sociales no relacionados con los contextos en los que se generaron o recolectaron originalmente los datos; (ii) Tratamiento perjudicial o desfavorable de ciertas personas naturales o grupos de personas que sea injustificado o desproporcionado a su comportamiento social o su gravedad.

4. **El uso de sistemas de identificación biométrica remota "en tiempo real" en espacios de acceso público con fines de aplicación de la ley, salvo excepciones tasadas y sujetas a autorización judicial:** (h) el uso de sistemas de identificación biométrica remota en tiempo real en espacios accesibles al público para fines de aplicación de la ley, a menos que y en la medida en que dicho uso sea estrictamente necesario para uno de los siguientes objetivos: (i) La búsqueda específica de víctimas de secuestro, trata de seres humanos o explotación sexual de seres humanos, así como la búsqueda de personas desaparecidas; (ii) La prevención de una amenaza específica, sustancial e inminente para la vida o la seguridad física de personas naturales o una amenaza genuina y presente o genuina y previsible de un ataque terrorista; (iii) La localización o identificación de una persona sospechosa de haber cometido un delito, con el propósito de llevar a cabo una investigación criminal o el enjuiciamiento o la ejecución de una pena criminal para los delitos referidos en el Anexo II y castigables en el Estado miembro correspondiente con una pena privativa de libertad o una orden de detención por un período máximo de al menos cuatro años.

Además, se establecen otras prohibiciones específicas:

5. **La comercialización, puesta en servicio o uso de sistemas de IA que creen o amplíen bases de datos de reconocimiento facial mediante la extracción no dirigida de imágenes faciales de Internet o grabaciones de CCTV:** (e) la comercialización, puesta en servicio para este propósito específico, o el uso de sistemas de IA que creen o amplíen bases de datos de reconocimiento facial mediante la extracción no dirigida de imágenes faciales de Internet o grabaciones de CCTV.
6. **El uso de sistemas de IA para inferir emociones de una persona natural en el ámbito laboral y en instituciones educativas, salvo cuando el uso del sistema de IA esté destinado a implementarse o comercializarse por razones médicas o de seguridad:** (f) la comercialización, puesta en servicio para este propósito específico, o el uso de sistemas de IA para inferir emociones de una persona natural en el ámbito laboral y en instituciones

educativas, excepto cuando el uso del sistema de IA esté destinado a implementarse o comercializarse por razones médicas o de seguridad.

7. **La comercialización, puesta en servicio o uso de sistemas de categorización biométrica que clasifiquen individualmente a las personas naturales en función de sus datos biométricos para deducir o inferir su raza, opiniones políticas, pertenencia a un sindicato, creencias religiosas o filosóficas, vida sexual u orientación sexual:** la comercialización, puesta en servicio para este propósito específico, o el uso de sistemas de categorización biométrica que clasifiquen individualmente a las personas naturales en función de sus datos biométricos para deducir o inferir su raza, opiniones políticas, pertenencia a un sindicato, creencias religiosas o filosóficas, vida sexual u orientación sexual; esta prohibición no cubre el etiquetado o filtrado de conjuntos de datos biométricos adquiridos legalmente, como imágenes, basados en datos biométricos o la categorización de datos biométricos en el área de la aplicación de la ley.

Estas prohibiciones reflejan la determinación del legislador europeo de establecer líneas rojas infranqueables en el desarrollo y despliegue de la IA, anteponiendo la dignidad y los derechos fundamentales de la persona a cualquier consideración de índole económica o tecnológica.

El antropocentrismo regulatorio se manifiesta también, de manera paradigmática, en la obligación de garantizar la supervisión humana en los sistemas de IA de alto riesgo (artículo 14). Esta disposición, que exige que los sistemas de IA de alto riesgo se diseñen y desarrollen de manera que puedan ser efectivamente supervisados por personas físicas, constituye un baluarte contra la autonomización descontrolada de los procesos decisarios basados en IA.

El artículo 14 especifica que la supervisión humana debe tener como objetivo prevenir o minimizar los riesgos para la salud, la seguridad o los derechos fundamentales, por ello, enumera una serie de medidas concretas que los proveedores y usuarios de sistemas de IA de alto riesgo deben implementar para garantizar una supervisión humana efectiva. Entre estas medidas se incluyen:

“Para la implementación de los párrafos 1, 2 y 3, el sistema de IA de alto riesgo se proporcionará al desplegador de manera que las personas naturales a quienes se les asigne la supervisión humana puedan, de manera adecuada y proporcional:

- (a) comprender adecuadamente las capacidades y limitaciones relevantes del sistema de IA de alto riesgo y ser capaces de monitorear debidamente su funcionamiento, incluido el objetivo de detectar y abordar anomalías, disfunciones y rendimientos inesperados;*
- (b) mantenerse conscientes de la posible tendencia a confiar automáticamente o sobreconfiar en el resultado producido por un sistema de IA de alto riesgo (sesgo de automatización), en particular para sistemas de IA de alto riesgo utilizados para proporcionar información o recomendaciones para decisiones a ser tomadas por personas naturales;*
- (c) interpretar correctamente el resultado del sistema de IA de alto riesgo, teniendo en cuenta, por ejemplo, las herramientas y métodos de interpretación disponibles;*
- (d) decidir, en cualquier situación particular, no utilizar el sistema de IA de alto riesgo o, de otro modo, desestimar, anular o revertir el resultado del sistema de IA de alto riesgo;*
- (e) intervenir en el funcionamiento del sistema de IA de alto riesgo o interrumpir el sistema mediante un botón de "detención" o un procedimiento similar que permita que el sistema se detenga en un estado seguro".*

Estas disposiciones reflejan la voluntad del legislador europeo de garantizar que, incluso en los sistemas más avanzados de IA, el ser humano mantenga siempre la última palabra en los procesos decisорios que afecten a derechos fundamentales.

b.- Regulación Basada en el Riesgo

El segundo principio cardinal que vertebría el AI Act es el de la regulación basada en el riesgo. Este enfoque, que modula la intensidad de la intervención normativa en función del nivel de riesgo que presenta cada aplicación específica de la IA, constituye una innovación jurídica de primer orden, que permite conciliar el fomento de la innovación tecnológica con la protección efectiva de los derechos fundamentales.

El reglamento establece una taxonomía del riesgo que distingue entre:

1. Sistemas de IA prohibidos (artículo 5).

2. Sistemas de IA de alto riesgo (artículos 6 y 7, y Anexo III).
3. Sistemas de IA sujetos a obligaciones de transparencia (artículo 52).
4. Sistemas de IA de riesgo mínimo o nulo.

Esta categorización, de una sofisticación y granularidad sin precedentes en el derecho regulatorio, permite una calibración precisa de las obligaciones normativas en función del riesgo potencial de cada sistema de IA.

Particular atención merece la categoría de sistemas de IA de alto riesgo, definida en el artículo 6 y desarrollada en el Anexo III. El reglamento considera de alto riesgo los sistemas de IA destinados a utilizarse como componentes de seguridad de productos sujetos a evaluación de conformidad *ex ante* por terceros, así como los sistemas de IA que se utilicen en una serie de ámbitos específicos enumerados en el Anexo III, entre los que se incluyen:

Los sistemas de IA de alto riesgo conforme al Artículo 6(2) son los sistemas de IA enumerados en cualquiera de las siguientes áreas:

Área	Descripción
Biometría	(a) Sistemas de identificación biométrica remota. No incluye los sistemas de IA destinados a la verificación biométrica para confirmar la identidad de una persona. (b) Sistemas de IA destinados a la categorización biométrica según atributos sensibles o protegidos basados en la inferencia de esos atributos. (c) Sistemas de IA destinados al reconocimiento de emociones.
Infraestructura crítica	Sistemas de IA destinados a ser utilizados como componentes de seguridad en la gestión y operación de infraestructura digital crítica, tráfico vial, o en el suministro de agua, gas, calefacción o electricidad.
Educación y formación profesional	(a) Sistemas de IA destinados a determinar el acceso o la admisión o para asignar a personas a instituciones educativas y de formación profesional. (b) Sistemas de IA destinados a evaluar los resultados del aprendizaje, incluyendo aquellos que guían el proceso de aprendizaje. (c) Sistemas de IA destinados a evaluar el nivel de educación que una persona recibirá o podrá acceder. (d) Sistemas de IA destinados a monitorear y detectar

	comportamientos prohibidos de los estudiantes durante pruebas.
Empleo, gestión de trabajadores y acceso al autoempleo	<p>(a) Sistemas de IA destinados al reclutamiento o selección de personas, incluyendo la colocación de anuncios de trabajo, análisis y filtrado de solicitudes, y evaluación de candidatos.</p> <p>(b) Sistemas de IA destinados a tomar decisiones sobre términos de relaciones laborales, promoción, terminación de relaciones contractuales, asignación de tareas basadas en comportamiento, rasgos personales o características, y monitoreo del rendimiento y comportamiento de los trabajadores.</p>
Acceso y disfrute de servicios privados esenciales y servicios y beneficios públicos esenciales	<p>(a) Sistemas de IA destinados a evaluar la elegibilidad de personas para beneficios y servicios públicos esenciales.</p> <p>(b) Sistemas de IA destinados a evaluar la solvencia crediticia de personas o establecer su puntaje de crédito, con excepción de los sistemas destinados a detectar fraudes financieros.</p> <p>(c) Sistemas de IA destinados a evaluar riesgos y precios en seguros de vida y salud.</p> <p>(d) Sistemas de IA destinados a evaluar y clasificar llamadas de emergencia o despachar servicios de respuesta.</p>
Aplicación de la ley	<p>(a) Sistemas de IA destinados a evaluar el riesgo de que una persona se convierta en víctima de delitos.</p> <p>(b) Sistemas de IA destinados a ser utilizados como polígrafo o herramientas similares.</p> <p>(c) Sistemas de IA destinados a evaluar la fiabilidad de pruebas en investigaciones o enjuiciamientos de delitos.</p> <p>(d) Sistemas de IA destinados a evaluar el riesgo de que una persona cometa delitos o reincida, basado en perfiles, rasgos de personalidad o comportamientos criminales.</p> <p>(e) Sistemas de IA destinados al perfil de personas según lo dispuesto en la Directiva (UE) 2016/680.</p>
Gestión de migración, asilo y control de fronteras	<p>(a) Sistemas de IA destinados a ser utilizados como polígrafo o herramientas similares.</p> <p>(b) Sistemas de IA destinados a evaluar riesgos, incluidos riesgos de seguridad, migración irregular, o salud.</p> <p>(c) Sistemas de IA destinados a asistir en el examen de solicitudes de asilo, visado o permisos de residencia.</p> <p>(d) Sistemas de IA destinados a detectar, reconocer o identificar personas naturales en el contexto de gestión</p>

	de migración, asilo o control de fronteras, excluyendo la verificación de documentos de viaje.
Administración de justicia y procesos democráticos	(a) Sistemas de IA destinados a asistir a una autoridad judicial en la investigación e interpretación de hechos y la ley, o en la resolución alternativa de disputas. (b) Sistemas de IA destinados a influir en el resultado de una elección o referéndum o en el comportamiento de votación de las personas, excluyendo los sistemas no directamente expuestos a las personas.

Esta lista, que el artículo 7 faculta a la comisión a ampliar mediante actos delegados, refleja la amplitud y profundidad del enfoque regulatorio europeo, que abarca, prácticamente, todos los ámbitos de la vida social y económica en los que la IA puede tener un impacto significativo en los derechos fundamentales.

Los sistemas de IA de alto riesgo están sujetos a una serie de requisitos obligatorios (artículos 8 a 15) que incluyen:

Tópico	Requisito
1. Establecimiento de un sistema de gestión de riesgos (Artículo 9)	<p>1.- Se establecerá, implementará, documentará y mantendrá un sistema de gestión de riesgos en relación con los sistemas de IA de alto riesgo.</p> <p>2.- El sistema será un proceso iterativo continuo planificado y ejecutado durante todo el ciclo de vida del sistema de IA, con revisión y actualización sistemática. Incluirá: (a) Identificación y análisis de riesgos conocidos y previsibles; (b) Estimación y evaluación de riesgos bajo condiciones de uso previstas y uso indebido; (c) Evaluación de otros riesgos según datos de monitoreo post-comercialización; (d) Adopción de medidas de gestión de riesgos específicas.</p> <p>3.- Los riesgos abordados serán aquellos mitigables o eliminables mediante diseño o información técnica.</p> <p>4.- Las medidas de gestión de riesgos considerarán las interacciones y efectos resultantes de la aplicación combinada de los requisitos.</p> <p>5.- Las medidas de gestión de riesgos garantizarán que los riesgos residuales sean aceptables, con acciones como eliminación de riesgos, implementación de medidas de mitigación y provisión de información técnica.</p> <p>6.- Los sistemas serán probados para identificar las medidas de gestión de riesgos más apropiadas.</p> <p>7.- Las pruebas pueden incluir pruebas en condiciones del mundo real.</p> <p>8.- Las pruebas se realizarán durante el desarrollo y antes</p>

	<p>de la comercialización o puesta en servicio.</p> <p>9.- Se considerará el impacto en menores de 18 años y otros grupos vulnerables.</p> <p>10.- Los procedimientos de gestión de riesgos podrán combinarse con otros procesos internos según la legislación de la Unión.</p>
2. Implementación de medidas de gobernanza de datos (Artículo 10)	<p>1.- Los sistemas de IA de alto riesgo que utilicen técnicas de entrenamiento de modelos de IA con datos deben basarse en conjuntos de datos que cumplan con los criterios de calidad.</p> <p>2.- Los conjuntos de datos estarán sujetos a prácticas de gobernanza y gestión de datos adecuadas para su propósito, incluyendo elecciones de diseño, procesos de recolección, preparación de datos, formulación de supuestos, evaluación de disponibilidad y adecuación, examen de sesgos, y medidas para prevenir y mitigar sesgos.</p> <p>3.- Los conjuntos de datos deben ser relevantes, representativos, libres de errores y completos para su propósito previsto.</p> <p>4.- Los conjuntos de datos deben considerar características geográficas, contextuales, conductuales o funcionales relevantes.</p> <p>5.- Los proveedores podrán procesar excepcionalmente categorías especiales de datos personales para corregir sesgos, siempre que se cumplan salvaguardas adecuadas.</p> <p>6.- Los párrafos 2 a 5 se aplicarán a los conjuntos de datos de prueba en sistemas de IA de alto riesgo que no utilicen técnicas de entrenamiento de modelos de IA.</p>
3. Elaboración de documentación técnica exhaustiva (Artículo 11)	<p>1.- La documentación técnica de un sistema de IA de alto riesgo se elaborará antes de su comercialización o puesta en servicio y se mantendrá actualizada. Deberá demostrar que el sistema cumple con los requisitos y proporcionar información clara y comprensible para la evaluación de conformidad. Contendrá los elementos del Anexo IV, pudiendo las PYME utilizar un formulario simplificado.</p> <p>2.- Cuando un sistema de IA de alto riesgo esté relacionado con un producto cubierto por la legislación de armonización de la Unión, se elaborará un único conjunto de documentación técnica que contenga toda la información necesaria.</p> <p>3.- La Comisión podrá modificar el Anexo IV para garantizar que la documentación técnica proporcione toda la información necesaria para evaluar la conformidad del sistema.</p>
4. Mantenimiento de registros automáticos (logs) (Artículo 12)	<p>1.- Los sistemas de IA de alto riesgo deberán permitir el registro automático de eventos durante toda su vida útil. Las</p>

	<p>capacidades de registro deberán permitir la grabación de eventos relevantes para: (a) Identificar situaciones de riesgo; (b) Facilitar el monitoreo post-comercialización; (c) Monitorear el funcionamiento del sistema.</p> <p>2.- Para ciertos sistemas de IA de alto riesgo, las capacidades de registro deberán incluir: (a) Registro del período de uso del sistema; (b) Base de datos de referencia; (c) Datos de entrada verificados; (d) Identificación de personas naturales involucradas en la verificación de resultados.</p>
5. Garantía de transparencia y suministro de información a los usuarios (Artículo 13)	<p>1.- Los sistemas de IA de alto riesgo deberán diseñarse con suficiente transparencia para permitir a los desplegadores interpretar y utilizar adecuadamente la salida del sistema. El tipo y grado de transparencia se garantizará para cumplir con las obligaciones del proveedor y el desplegador.</p> <p>2.- Los sistemas de IA de alto riesgo deberán ir acompañados de instrucciones de uso en formato digital o de otra forma accesible y comprensible para los desplegadores.</p> <p>3.- Las instrucciones de uso deberán contener al menos: (a) Identidad y datos de contacto del proveedor; (b) Características, capacidades y limitaciones de rendimiento del sistema; (c) Cambios predeterminados en el sistema; (d) Medidas de supervisión humana; (e) Recursos computacionales y de hardware necesarios; (f) Descripción de mecanismos para recolectar y almacenar registros.</p>
6. Implementación de medidas de supervisión humana (Artículo 14)	<p>1.- Los sistemas de IA de alto riesgo deberán diseñarse con herramientas de interfaz humano-máquina para ser supervisados eficazmente por personas naturales durante su uso.</p> <p>2.- La supervisión humana deberá minimizar los riesgos para la salud, seguridad o derechos fundamentales que puedan surgir durante el uso del sistema.</p> <p>3.- Las medidas de supervisión deberán ser proporcionales a los riesgos y al contexto de uso, e incluir medidas integradas en el sistema o implementadas por el desplegador.</p> <p>4.- El sistema deberá permitir a los supervisores humanos comprender sus capacidades, monitorear su operación, y tomar decisiones adecuadas.</p> <p>5.- En ciertos sistemas de IA de alto riesgo, se requerirá la verificación de la identificación por al menos dos personas naturales antes de tomar decisiones basadas en esa identificación, excepto en casos de aplicación de la ley, migración, control fronterizo o asilo.</p>

7. Cumplimiento de requisitos de precisión, robustez y ciberseguridad (Artículo 15)	<p>1.- Los sistemas de IA de alto riesgo deberán diseñarse para lograr un nivel adecuado de precisión, robustez y ciberseguridad durante todo su ciclo de vida.</p> <p>2.- La Comisión fomentará el desarrollo de puntos de referencia y metodologías de medición para niveles apropiados de precisión y robustez.</p> <p>3.- Los niveles de precisión deberán declararse en las instrucciones de uso.</p> <p>4.- Los sistemas de IA de alto riesgo deberán ser resistentes a errores, fallos o inconsistencias, y se tomarán medidas técnicas y organizativas para lograrlo, incluyendo soluciones de redundancia técnica y planes de respaldo.</p> <p>5.- Los sistemas de IA de alto riesgo deberán ser resistentes frente a intentos de terceros no autorizados de alterar su uso, salidas o rendimiento mediante la explotación de vulnerabilidades del sistema. Se implementarán medidas técnicas adecuadas para garantizar la ciberseguridad.</p>
--	---

Estos requisitos configuran un marco normativo que busca garantizar que los sistemas de IA de alto riesgo se desarrollem y desplieguen de manera segura, transparente y respetuosa con los derechos fundamentales.

C. Transparencia y Trazabilidad

El tercer principio fundamental que informa el AI Act es el de transparencia y trazabilidad. Este principio, que se materializa en una serie de obligaciones concretas para los proveedores y usuarios de sistemas de IA, busca disipar la opacidad que, tradicionalmente, ha caracterizado a los algoritmos de IA, facilitando su supervisión regulatoria y empoderando a los usuarios finales.

El artículo 13 del Reglamento impone a los proveedores de sistemas de IA de alto riesgo la obligación de garantizar que estos sistemas se diseñen y desarrollen de manera que sean suficientemente transparentes para permitirles a los usuarios interpretar y utilizar adecuadamente sus resultados. Esta obligación se concreta en el deber de acompañar los sistemas de IA de alto riesgo con instrucciones de uso que incluyan:

1. Las características, capacidades y limitaciones de rendimiento del sistema.
2. Los cambios predeterminados por el proveedor en el sistema y su rendimiento.

3. Las medidas de supervisión humana.
4. La vida útil esperada del sistema y las medidas de mantenimiento necesarias.

Estas disposiciones buscan garantizar que los usuarios de sistemas de IA de alto riesgo dispongan de toda la información necesaria para utilizarlos de manera segura y responsable.

Por su parte, el artículo 12 establece la obligación de que los sistemas de IA de alto riesgo sean capaces de registrar automáticamente los eventos (logs) a lo largo de su ciclo de vida. Esta obligación de trazabilidad, permitirá una supervisión y auditoría exhaustiva del funcionamiento de los sistemas de IA de alto riesgo, facilitando la detección y corrección de posibles sesgos o errores.

D.- Gobernanza Multinivel

El cuarto principio rector del AI Act es el de gobernanza multinivel. El reglamento establece un sofisticado sistema de gobernanza que involucra a instituciones europeas, autoridades nacionales y organismos notificados, garantizando así una implementación coherente y efectiva del marco regulatorio en todo el territorio de la Unión. Este enfoque refleja la complejidad inherente a la regulación de la IA y la necesidad de una coordinación estrecha entre diferentes niveles de gobierno y áreas de expertise. Los principales componentes de este sistema de gobernanza son:

1. AI Office (Oficina de IA).

El AI Act establece la creación de una Oficina de IA a nivel de la Unión Europea, como se detalla en el artículo 64. Esta oficina, integrada en la estructura de la Comisión Europea, tiene como misión desarrollar la experiencia y las capacidades de la Unión en el campo de la IA.

2. European Artificial Intelligence Board (Junta Europea de Inteligencia Artificial).

El artículo 65 establece la creación de la Junta Europea de Inteligencia Artificial. Este organismo, compuesto por representantes de los Estados miembros y de la Comisión, tiene un papel crucial en la coordinación y cooperación entre las autoridades nacionales y la Comisión. Sus responsabilidades incluyen (Artículo 66):

“(a) contribuir a la coordinación entre las autoridades nacionales competentes responsables de la aplicación de este Reglamento y, en cooperación y sujeto al acuerdo de las autoridades de vigilancia del mercado, apoyar las actividades conjuntas de las autoridades de vigilancia del mercado mencionadas en el Artículo 74(11);

(b) recopilar y compartir conocimientos técnicos y normativos y mejores prácticas entre los Estados miembros;

(c) proporcionar asesoramiento sobre la implementación de este Reglamento, en particular en lo que respecta a la aplicación de normas sobre modelos de IA de propósito general;

(d) contribuir a la armonización de prácticas administrativas en los Estados miembros, incluidas las relacionadas con la exención de los procedimientos de evaluación de conformidad mencionados en el Artículo 46, el funcionamiento de los entornos regulatorios de IA y las pruebas en condiciones del mundo real mencionadas en los Artículos 57, 59 y 60;

(e) a solicitud de la Comisión o por iniciativa propia, emitir recomendaciones y opiniones escritas sobre cualquier asunto relevante relacionado con la implementación de este Reglamento y su aplicación coherente y efectiva, incluyendo:

(i) sobre el desarrollo y aplicación de códigos de conducta y códigos de práctica conforme a este Reglamento, así como las directrices de la Comisión;

(ii) la evaluación y revisión de este Reglamento conforme al Artículo 112, incluidos los informes de incidentes graves mencionados en el Artículo 73, y el funcionamiento de la base de datos de la UE mencionada en el Artículo 71, la preparación de los actos delegados o de ejecución, y en lo que respecta a posibles alineaciones de este Reglamento con la legislación de armonización de la Unión enumerada en el Anexo I;

(iii) sobre especificaciones técnicas o normas existentes en relación con los requisitos establecidos en el Capítulo III, Sección 2;

- (iv) sobre el uso de normas armonizadas o especificaciones comunes mencionadas en los Artículos 40 y 41;
 - (v) tendencias, como la competitividad global europea en IA, la adopción de IA en la Unión y el desarrollo de habilidades digitales;
 - (vi) tendencias en la tipología evolutiva de las cadenas de valor de IA, en particular sobre las implicaciones resultantes en términos de responsabilidad;
 - (vii) sobre la posible necesidad de enmienda del Anexo III de conformidad con el Artículo 7, y sobre la posible necesidad de revisión del Artículo 5 conforme al Artículo 112, teniendo en cuenta la evidencia disponible relevante y los últimos desarrollos tecnológicos;
- f) apoyar a la Comisión en la promoción de la alfabetización en IA, la concienciación pública y la comprensión de los beneficios, riesgos, salvaguardias y derechos y obligaciones en relación con el uso de sistemas de IA;
- (g) facilitar el desarrollo de criterios comunes y una comprensión compartida entre los operadores del mercado y las autoridades competentes de los conceptos relevantes establecidos en este Reglamento, incluyendo la contribución al desarrollo de puntos de referencia;
- (h) cooperar, según sea apropiado, con otras instituciones, organismos, oficinas y agencias de la Unión, así como con grupos y redes de expertos relevantes de la Unión, en particular en los campos de la seguridad de productos, ciberseguridad, competencia, servicios digitales y de medios, servicios financieros, protección del consumidor, protección de datos y derechos fundamentales;
- (i) contribuir a una cooperación efectiva con las autoridades competentes de terceros países y con organizaciones internacionales;
- (j) asistir a las autoridades nacionales competentes y a la Comisión en el desarrollo de la experiencia organizativa y técnica necesaria para la implementación de este Reglamento,

incluyendo la contribución a la evaluación de las necesidades de capacitación del personal de los Estados miembros involucrados en la implementación de este Reglamento;

(k) asistir a la Oficina de IA en el apoyo a las autoridades nacionales competentes en el establecimiento y desarrollo de entornos regulatorios de IA y facilitar la cooperación y el intercambio de información entre los entornos regulatorios de IA;

(l) contribuir y proporcionar asesoramiento relevante sobre el desarrollo de documentos de orientación;

(m) asesorar a la Comisión en relación con asuntos internacionales sobre IA;

(n) proporcionar opiniones a la Comisión sobre las alertas calificadas respecto a los modelos de IA de propósito general;

(o) recibir opiniones de los Estados miembros sobre las alertas calificadas respecto a los modelos de IA de propósito general, y sobre experiencias y prácticas nacionales en la supervisión y aplicación de sistemas de IA, en particular sistemas que integran los modelos de IA de propósito general”.

La creación de este organismo supranacional refleja la voluntad del legislador europeo de garantizar una aplicación uniforme del reglamento en todo el territorio de la Unión, evitando así la fragmentación regulatoria que podría socavar la eficacia del marco normativo.

3. Advisory Forum (Foro Consultivo).

El artículo 67 prevé la creación de un Foro Consultivo³. Este organismo está diseñado para proporcionar experiencia técnica y asesoramiento a la junta y a la comisión. Su composición refleja un enfoque equilibrado, incluyendo:

- a) Representantes de la industria
- b) Startups y PYMEs
- c) Sociedad civil

d) Academia

La membresía del foro asesor debe equilibrada en cuanto a intereses comerciales y no comerciales y, dentro de la categoría de intereses comerciales, en cuanto a pymes y otras empresas. La inclusión de un foro consultivo en la estructura de gobernanza subraya el compromiso del legislador europeo con un enfoque participativo en la regulación de la IA. En el contexto de la administración de justicia, este foro podría desempeñar un papel crucial en la identificación de las implicaciones éticas y sociales de la implementación de sistemas de IA en los procesos judiciales.

4. Scientific Panel of Independent Experts (Panel Científico de Expertos Independientes).

El artículo 68 establece la creación de un Panel Científico de Expertos Independientes⁴. Este panel tiene como objetivo proporcionar asesoramiento científico imparcial a la AI Office. Sus funciones incluyen:

“(a) apoyar la implementación y aplicación de este Reglamento en lo que respecta a los modelos y sistemas de IA de propósito general, en particular:

(i) alertar a la Oficina de IA sobre posibles riesgos sistémicos a nivel de la Unión de modelos de IA de propósito general, de conformidad con el Artículo 90;

(ii) contribuir al desarrollo de herramientas y metodologías para evaluar las capacidades de los modelos y sistemas de IA de propósito general, incluidos los puntos de referencia;

(iii) proporcionar asesoramiento sobre la clasificación de modelos de IA de propósito general con riesgo sistémico;

(iv) proporcionar asesoramiento sobre la clasificación de diversos modelos y sistemas de IA de propósito general;

(v) contribuir al desarrollo de herramientas y plantillas;

(b) apoyar el trabajo de las autoridades de vigilancia del mercado, a solicitud de estas;

(c) apoyar las actividades de vigilancia del mercado transfronterizas mencionadas en el Artículo 74(11), sin perjuicio de los poderes de las autoridades de vigilancia del mercado;

d) apoyar a la Oficina de IA en el desempeño de sus funciones en el contexto del procedimiento de salvaguardia de la Unión conforme al Artículo 81.”

La inclusión de un panel científico independiente en la estructura de gobernanza es particularmente relevante en el contexto de la administración de justicia, donde la evaluación rigurosa e imparcial de los sistemas de IA es crucial para mantener la integridad del proceso judicial.

5. Autoridades Nacionales Competentes.

A nivel nacional, el artículo 69 obliga a los Estados miembros a designar una o varias autoridades nacionales competentes para supervisar la aplicación y ejecución del reglamento¹⁰. Estas autoridades incluyen, al menos, una autoridad notificadora y una autoridad de vigilancia del mercado. Su función principal es ejercer sus poderes de manera independiente y sin sesgo para salvaguardar la objetividad de sus actividades y tareas. Además, deben disponer de recursos técnicos, financieros y humanos adecuados, incluyendo personal con experiencia en IA, protección de datos, ciberseguridad y derechos fundamentales. También deben asegurar un nivel adecuado de ciberseguridad y actuar conforme con las obligaciones de confidencialidad establecidas.

Las autoridades nacionales competentes tienen la tarea de proporcionar orientación y asesoramiento sobre la implementación del reglamento, especialmente a pymes y start-ups, en coordinación con la junta y la comisión. Además, deben facilitar la coherencia y coordinación entre las autoridades competentes nacionales, incluyendo la recopilación de datos relevantes. Deben informar regularmente a la Comisión sobre el estado de sus recursos y facilitar el intercambio de experiencias entre autoridades nacionales. Para las instituciones de la Unión, el Supervisor Europeo de Protección de Datos actúa como la autoridad competente para su supervisión.

6. Autoridad Notificadora.

El sistema de gobernanza multinivel se completa con las figuras de la autoridad notificadora y organismos notificados, reguladas en los artículos 28 a 39 del AI Act. Las autoridades notificadoras, designadas por los Estados miembros de la UE, son entidades encargadas de establecer y ejecutar los procedimientos necesarios para la evaluación, designación y notificación de los organismos de evaluación de la conformidad, así como de su monitoreo continuo. Estas autoridades desempeñan un papel crucial en la supervisión y garantía de la integridad del sistema de evaluación de conformidad. El Artículo 28 del EU AI Act establece que las autoridades notificadoras deben cumplir con requisitos estrictos, incluyendo la salvaguarda de la objetividad e imparcialidad de sus actividades, la adopción de decisiones por personas competentes distintas de las que realizaron la evaluación, la prohibición de ofrecer servicios comerciales o de consultoría que puedan entrar en conflicto con sus funciones, la preservación de la confidencialidad de la información obtenida y la disposición de personal competente suficiente para el desempeño adecuado de sus tareas.

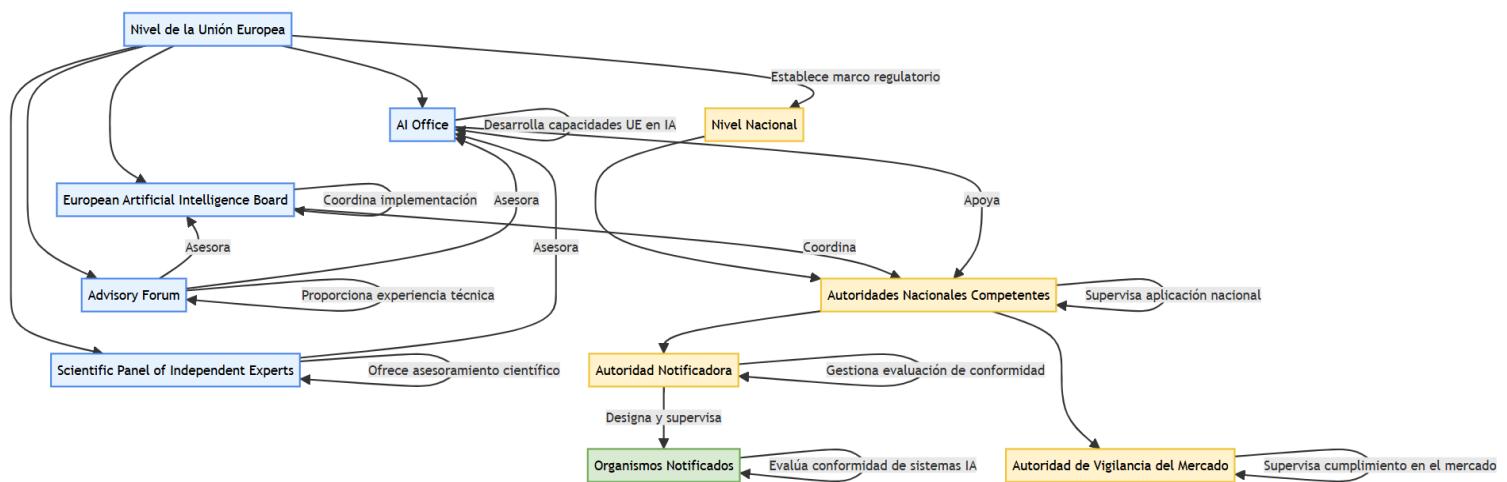
Por su parte, los organismos notificados, como se define en el Artículo 31 del EU AI Act, son entidades de evaluación de la conformidad designadas por las autoridades notificadoras para llevar a cabo las tareas de evaluación de la conformidad de los sistemas de IA de alto riesgo. Estos organismos juegan un papel fundamental en la verificación del cumplimiento de los requisitos establecidos en el EU AI Act. Para ser designado como organismo notificado, una entidad debe cumplir con requisitos rigurosos, incluyendo estar establecida de conformidad con el Derecho nacional de un Estado miembro y tener personalidad jurídica, ser independiente de la organización o el sistema de IA de alto riesgo que evalúa, contar con los procedimientos y la experiencia necesarios para llevar a cabo las actividades de evaluación de la conformidad y, además, demostrar un alto nivel de competencia técnica, profesional y científica en el campo de la IA.

El proceso de notificación, detallado en el Artículo 30 del EU AI Act, implica una solicitud detallada a la autoridad notificadora, que evalúa la competencia del organismo. Una vez aprobado, se notifica a la Comisión Europea y a los demás Estados miembros. Los organismos notificados tienen obligaciones operativas específicas, incluyendo llevar a cabo evaluaciones de conformidad con un alto grado de profesionalidad e integridad técnica, tener en cuenta el tamaño de la empresa,

el sector en que opera, su estructura y el grado de complejidad del sistema de IA en cuestión, además de participar en actividades de coordinación y normalización pertinentes.

La relación entre las autoridades notificadoras y los organismos notificados es de supervisión continua. Las autoridades notificadoras tienen la facultad, según el Artículo 36 del EU AI Act, de restringir, suspender o retirar la designación de un organismo notificado si este deja de cumplir los requisitos o de desempeñar sus obligaciones. Este sistema de control mutuo asegura la integridad y eficacia del proceso de evaluación de conformidad.

El sistema de autoridades notificadoras y organismos notificados establecido por el EU AI Act es crucial para la implementación efectiva de la regulación de la IA en la Unión Europea. Proporciona un mecanismo de control de calidad independiente y técnicamente competente para evaluar la conformidad de los sistemas de IA de alto riesgo con los requisitos legales. Este sistema contribuye a garantizar la seguridad, la fiabilidad y el respeto de los derechos fundamentales en el desarrollo y uso de la IA en la UE, al tiempo que facilita la innovación y el desarrollo del mercado único digital.



E. Extraterritorialidad y Efecto Bruselas

El principio de extraterritorialidad en el AI Act representa una manifestación paradigmática de la creciente influencia normativa de la Unión Europea en el ámbito global, fenómeno que ha

sido acuñado por la doctrina como "efecto Bruselas"¹⁹⁴. Este principio, consagrado en el artículo 2 del reglamento, trasciende la concepción tradicional de la jurisdicción territorial para proyectar los efectos normativos del AI Act más allá de las fronteras de la Unión.

La vocación extraterritorial del AI Act se articula en torno a tres ejes fundamentales:

1. **Aplicación a proveedores de terceros países:** el artículo 2(1)(c) establece que el Reglamento se aplica a los "*proveedores que comercialicen o pongan en servicio sistemas de IA en la Unión, independientemente de si dichos proveedores están establecidos en la Unión o en un tercer país*". Esta disposición extiende el ámbito de aplicación del Reglamento a todas las empresas que deseen acceder al mercado único europeo, independientemente de su lugar de establecimiento.
2. **Aplicación a usuarios de la UE de sistemas de IA de terceros países:** el artículo 2(1)(b) prevé la aplicación del reglamento a los "*usuarios de sistemas de IA ubicados en la Unión*"³. Esta disposición, interpretada en conjunción con el artículo 2(1)(c), implica que incluso los sistemas de IA desarrollados y operados enteramente fuera de la UE caerán bajo el ámbito de aplicación del reglamento si sus outputs son utilizados por usuarios ubicados en la Unión.
3. **Aplicación basada en el impacto en la UE:** el artículo 2(1)(c) extiende la aplicación del Reglamento a los "*proveedores y usuarios de sistemas de IA que estén ubicados en un tercer país, cuando la producción generada por el sistema se utilice en la Unión*"⁴. Esta disposición, de una amplitud palmaria, permite la aplicación del reglamento a operadores de terceros países que ni siquiera tienen una presencia física en la UE, basándose únicamente en el impacto de sus sistemas en el territorio de la Unión.

La justificación jurídica de esta extraterritorialidad se fundamenta en la doctrina de los "efectos", desarrollada por el Tribunal de Justicia de la UE en casos como Woodpulp¹⁹⁵ y Intel¹⁹⁶.

¹⁹⁴ Bradford, Anu. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press, 2020, p. 25-67. Recuperado de: <https://dokumen.pub/qdownload/the-brussels-effect-how-the-european-union-rules-the-world-9780190088583-2019031328-2019031329-9780190088606-9780190088590-9780190088613.html>

¹⁹⁵ Andrew N. Vollmer & John Byron Sandage, "The Wood Pulp Case," 23 *Int'l L.* 721 (1989), <https://scholar.smu.edu/tl/vol23/iss3/9>

¹⁹⁶ Tribunal de Justicia de la Unión Europea (Tribunal General, Sala Séptima, Composición Ampliada), 12 de junio de 2014, Intel Corp. contra Comisión Europea, asunto T-286/09, ECLI:EU:T:2014:547, <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62009TJ0286>

Según esta doctrina, la UE puede aplicar su legislación a conductas que ocurren fuera de su territorio si dichas conductas producen efectos dentro de la Unión.

El "efecto Bruselas" que se espera que produzca el AI Act se puede analizar desde tres perspectivas:

1. **Efecto de jure:** se refiere a la adopción formal de estándares similares a los de la UE por parte de terceros países. En el caso del AI Act, esto podría manifestarse en la adopción de legislaciones nacionales inspiradas en el reglamento europeo, como ya ha ocurrido con el RGPD¹⁹⁷.
2. **Efecto de facto:** implica la adopción voluntaria de los estándares de la UE por parte de empresas globales, con el fin de mantener el acceso al mercado único europeo. Este efecto ya se ha observado con el RGPD, donde numerosas empresas tecnológicas han optado por aplicar los estándares europeos de protección de datos a nivel global¹⁹⁸.
3. **Efecto de mercado:** se refiere a la influencia que la regulación de la UE puede tener en los estándares de facto del mercado global. En el caso del AI Act, esto podría manifestarse en la adopción generalizada de prácticas de desarrollo y despliegue de IA alineadas con los requisitos del reglamento, incluso en jurisdicciones donde no existe una obligación legal de hacerlo⁹.

En el contexto específico de la administración de justicia, la extraterritorialidad del AI Act podría tener implicaciones significativas. Por ejemplo:

- Los proveedores de sistemas de IA judicial establecidos fuera de la UE que deseen ofrecer sus servicios a tribunales europeos deberán cumplir con los requisitos del reglamento,
- Los tribunales de la UE que utilicen sistemas de IA desarrollados en terceros países deberán asegurarse de que dichos sistemas cumplan con el reglamento,

¹⁹⁷ Graham Greenleaf, "Global Data Privacy Laws 2021: Despite COVID Delays, 145 Laws Show GDPR Dominance," *Privacy Laws & Business International Report*, no. 169 (2021): 1-5. Recuperado de: [ssrn_id3933588_code722134.pdf \(elsevier-ssrn-document-store-prod.s3.amazonaws.com\)](https://ssrn.com/abstract=3933588)

¹⁹⁸ Bilyana Petkova, "Privacy as Europe's First Amendment," *European Law Journal* 25, no. 2 (2019): 140-154. Recuperado de: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333937

- Las decisiones judiciales de la UE basadas en outputs de sistemas de IA podrían ser impugnadas si dichos sistemas no cumplen con el reglamento, incluso si fueron desarrollados y operados fuera de la Unión.

Sin embargo, la aplicación extraterritorial del AI Act también plantea desafíos significativos:

1. **Conflictos de jurisdicción:** podrían surgir conflictos con las leyes de terceros países que regulen los mismos sistemas de IA de manera diferente.
2. **Dificultades de *enforcement*:** la aplicación efectiva del reglamento a operadores de terceros países podría resultar compleja en ausencia de acuerdos de cooperación internacional.
3. **Possibles represalias:** la extraterritorialidad del AI Act podría ser percibida por algunos países como una forma de imperialismo regulatorio, lo que podría llevar a medidas de represalia.

Concluyendo, la extraterritorialidad del AI Act representa una apuesta ambiciosa del legislador europeo por establecer un estándar global en la regulación de la IA. Si bien este enfoque plantea desafíos significativos, también ofrece la oportunidad de promover un desarrollo de la IA alineado con los valores fundamentales de la UE a escala global.

F. Principio de Precaución

El principio de precaución, consagrado en el artículo 191 del Tratado de Funcionamiento de la Unión Europea¹⁹⁹, se erige como uno de los pilares fundamentales del AI Act. Este principio, que ha sido desarrollado extensamente por la jurisprudencia del Tribunal de Justicia de la UE²⁰⁰, justifica la adopción de medidas protectoras sin tener que esperar a que se demuestre plenamente la realidad y gravedad de tales riesgos.

¹⁹⁹ Versión Consolidada del Tratado de Funcionamiento de la Unión Europea [2012]. Recuperado de: [Versión consolidada del Tratado de Funcionamiento de la Unión Europea \(boe.es\)](https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A61999TJ0013)

²⁰⁰ Sentencia del Tribunal de Primera Instancia de 11 de septiembre de 2002, Pfizer Animal Health SA contra Consejo, T-13/99, ECLI:EU:T:2002:209. Recuperada de: <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A61999TJ0013>

En el contexto del AI Act, el principio de precaución se manifiesta de diversas formas:

1. **Prohibiciones absolutas:** el artículo 5 del Reglamento establece una serie de prácticas de IA prohibidas, incluyendo: a) El uso de técnicas subliminales (Art. 5.1.a) b) La explotación de vulnerabilidades de grupos específicos (Art. 5.1.b) c) El uso de sistemas de puntuación social (Art. 5.1.c) d) El uso de sistemas de identificación biométrica remota "en tiempo real" en espacios de acceso público con fines de aplicación de la ley, salvo excepciones tasadas (Art. 5.1.d). Estas prohibiciones reflejan la aplicación más estricta del principio de precaución, al vetar completamente prácticas consideradas de riesgo inaceptable.
2. **Enfoque basado en el riesgo:** el reglamento adopta un enfoque graduado basado en el riesgo, clasificando los sistemas de IA en categorías de riesgo y estableciendo requisitos proporcionales para cada categoría. Este enfoque, que se plasma en los artículos 6 y 7 y en el Anexo III, permite una aplicación matizada del principio de precaución, calibrando la intensidad de la intervención regulatoria en función del nivel de riesgo percibido.
3. **Requisitos para sistemas de IA de alto riesgo:** los artículos 8 a 15 establecen una serie de requisitos rigurosos para los sistemas de IA de alto riesgo, incluyendo: a) Sistema de gestión de riesgos (Art. 9) b) Gobernanza de datos (Art. 10) c) Documentación técnica (Art. 11) d) Registro de eventos (Art. 12) e) Transparencia (Art. 13) f) Supervisión humana (Art. 14) g) Precisión, robustez y ciberseguridad (Art. 15). Estos requisitos reflejan la aplicación del principio de precaución a través de la imposición de salvaguardias técnicas y organizativas exhaustivas.
4. **Mecanismos de adaptación dinámica:** el artículo 7 faculta a la Comisión para ampliar la lista de sistemas de IA de alto riesgo mediante actos delegados¹⁵. Esta disposición permite una aplicación dinámica del principio de precaución, adaptando el marco regulatorio a la evolución tecnológica y a la aparición de nuevos riesgos.

En el contexto específico de la administración de justicia, el principio de precaución adquiere una relevancia particular:

1. **Clasificación como alto riesgo:** el Anexo III, punto 8, clasifica como de alto riesgo los sistemas de IA destinados a asistir a una autoridad judicial en la investigación e interpretación de hechos y del Derecho y en la aplicación de la ley a un conjunto concreto

de hechos¹⁶. Esta clasificación refleja una aplicación del principio de precaución, sometiendo estos sistemas a los requisitos más estrictos del reglamento.

2. **Supervisión humana:** el artículo 14, al exigir medidas de supervisión humana para los sistemas de IA de alto riesgo, refleja una aplicación del principio de precaución en el ámbito judicial. Esta disposición busca preservar la autonomía decisoria del juez y prevenir la automatización descontrolada de los procesos judiciales¹⁷.
3. **Transparencia y explicabilidad:** los requisitos de transparencia y explicabilidad establecidos en el artículo 13 pueden interpretarse como una manifestación del principio de precaución en el contexto judicial. Estas disposiciones buscan garantizar que las decisiones judiciales basadas en outputs de sistemas de IA sean comprensibles y susceptibles de escrutinio¹⁸.

La aplicación del principio de precaución en el AI Act, si bien busca prevenir daños potencialmente irreversibles a los derechos fundamentales y al tejido social, no está exenta de críticas:

1. **Potencial freno a la innovación:** un enfoque excesivamente precautorio podría inhibir el desarrollo y despliegue de sistemas de IA beneficiosos.
2. **Subjetividad en la evaluación del riesgo:** la determinación de qué constituye un riesgo "inaceptable" o "alto" puede ser subjetiva y estar sujeta a sesgos.
3. **Possible desventaja competitiva:** un marco regulatorio excesivamente restrictivo podría poner en desventaja a las empresas europeas frente a competidores de jurisdicciones menos reguladas.

Sobre este último punto, el artículo "Europe and AI: Causes and Implications of Europe Losing Ground in the Race for AI (Part I)"²⁰¹ de Richard Oubělický resulta particularmente ilustrativo, pues analiza la posición de Europa en la carrera global por el desarrollo de la inteligencia artificial (IA), centrándose en las causas y posibles implicaciones del aparente rezago europeo en este campo.

²⁰¹ Richard Oubělický, "Europe and AI: Causes and Implications of Europe Losing Ground in the Race for AI (Part I)," Security Outlines, 13 de marzo de 2024, <https://securityoutlines.cz/europe-and-ai-causes-and-implications-of-europe-losing-ground-in-the-race-for-ai-part-i/>

El autor comienza destacando la creciente importancia de la IA y su impacto económico proyectado: *"Por ejemplo, para 2030, se espera que hasta el 45% de los ingresos globales estén vinculados al uso de la IA."* Esta proyección subraya la relevancia estratégica de la IA y la necesidad de que Europa mantenga su competitividad en este campo.

En cuanto a la inversión en IA, el texto señala una disparidad significativa: *"En los últimos años, Estados Unidos y China representan aproximadamente el 75-80% de la inversión global en IA (76,6% - 2022, 76% - 2021, 80% - 2020)."* Esta concentración de inversiones en dos países plantea desafíos para la competitividad europea.

A partir de lo anterior, Oubělický contrasta los enfoques de China, Estados Unidos y la Unión Europea en el desarrollo y regulación de la IA, centrando su exposición sobre el argumento de que la implementación de las normas regulatorias más precavidas en el EU AI Act, aunque diseñadas para mitigar riesgos potenciales, podrían resultar en una carga regulatoria desproporcionada.

La implementación de estas medidas regulatorias conlleva costos de cumplimiento sustanciales que podrían desviar recursos críticos de las actividades de investigación y desarrollo, mermando así la capacidad innovadora de las empresas europeas. Además, los procesos de evaluación previa y obtención de autorizaciones regulatorias podrían introducir demoras significativas en el ciclo de desarrollo y comercialización de soluciones de IA, poniendo a las empresas europeas en desventaja frente a competidores de jurisdicciones con marcos regulatorios más flexibles.

El artículo subraya la preocupación de que un entorno regulatorio percibido como excesivamente restrictivo podría desincentivar la atracción de capital humano y financiero hacia el ecosistema europeo de IA. Esta fuga de talento e inversión hacia jurisdicciones percibidas como más favorables a la innovación podría exacerbar la brecha tecnológica ya existente entre la UE y otros líderes globales en IA como Estados Unidos y China.

La asimetría regulatoria global que se deriva de los enfoques divergentes adoptados por las principales potencias en IA plantea desafíos significativos para la competitividad europea. Mientras que Estados Unidos ha optado por un enfoque más flexible basado en directrices

sectoriales y autorregulación industrial y China ha adoptado una estrategia que combina el fomento activo del desarrollo de IA con controles gubernamentales específicos, la UE se encuentra en riesgo de imponer un marco regulatorio que, aunque bien intencionado, podría resultar excesivamente oneroso para sus empresas.

In fine, el principio de precaución, pilar fundamental del AI Act, se manifiesta de diversas formas en la regulación de la inteligencia artificial en la Unión Europea, desde prohibiciones absolutas hasta requisitos rigurosos para sistemas de alto riesgo. Este enfoque, si bien busca proteger los derechos fundamentales y prevenir daños potencialmente irreversibles, plantea importantes desafíos y suscita críticas significativas.

La aplicación del principio de precaución en el contexto de la IA, particularmente en ámbitos sensibles como la administración de justicia, refleja la voluntad del legislador europeo de anteponer la seguridad y la protección de derechos a la innovación descontrolada. Sin embargo, como señala el análisis de Oubělický, este enfoque podría tener consecuencias no deseadas en términos de competitividad global.

La tensión entre la necesidad de regular para proteger y el imperativo de innovar para competir se erige como uno de los principales dilemas que enfrenta la Unión Europea en el campo de la IA. El riesgo de que un marco regulatorio excesivamente restrictivo pueda desincentivar la inversión, ralentizar la innovación y provocar una fuga de talento hacia jurisdicciones más permisivas no puede ser subestimado.

En última instancia, el éxito del AI Act y, por extensión, de la estrategia europea en materia de IA, dependerá de su capacidad para encontrar un equilibrio adecuado entre precaución e innovación. Esto requerirá una aplicación flexible y adaptativa del principio de precaución, que permita salvaguardar los valores fundamentales de la UE sin comprometer su competitividad en un sector tecnológico de importancia estratégica crítica.

La búsqueda de este equilibrio no solo es crucial para el futuro tecnológico de Europa, sino que también podría sentar las bases para un modelo global de regulación de la IA que armonice el progreso tecnológico con la protección de los derechos humanos y los valores democráticos.

G. Responsabilidad y Rendición de Cuentas

En el contexto del desarrollo y uso de sistemas de Inteligencia Artificial (IA), la responsabilidad y la rendición de cuentas son principios fundamentales para garantizar que estas tecnologías se utilicen de manera ética, segura y confiable. La IA tiene el potencial de tomar decisiones y realizar acciones que pueden tener un impacto significativo en las personas, la sociedad y el ambiente. Por lo tanto, es crucial establecer mecanismos claros para determinar quién es responsable de las consecuencias de estas decisiones y acciones y cómo se puede exigir que rindan cuentas.

El Acta de IA de la UE reconoce la importancia de estos principios y establece varias medidas para promoverlos. Estas incluyen obligaciones específicas para los proveedores y usuarios de sistemas de IA, los requisitos de transparencia y supervisión humana y la creación de un marco para el registro y monitoreo de sistemas de alto riesgo.

1. Obligaciones de los Proveedores.

El Acta impone una serie de obligaciones a los proveedores de sistemas de IA, especialmente aquellos considerados de alto riesgo. Estas obligaciones incluyen:

- Establecer y mantener un sistema de gestión de calidad que asegure el cumplimiento de los requisitos del Acta (Art. 17),
- Llevar a cabo una evaluación de conformidad antes de comercializar o poner en servicio un sistema de alto riesgo (Art. 43),
- Registrar los sistemas de alto riesgo en una base de datos de la UE (Art. 49),
- Monitorear y reportar incidentes graves y mal funcionamientos (Art. 73).

Lo anterior busca garantizar que los proveedores asuman la responsabilidad de la seguridad y conformidad de sus sistemas, además de que exista un rastro claro de responsabilidad en caso de problemas.

2. Registro de Sistemas de Alto Riesgo.

El acta crea una base de datos a nivel de la UE para registrar los sistemas de IA de alto riesgo (Art. 49). Los proveedores deben registrar sus sistemas antes de comercializarlos o ponerlos

en servicio, proporcionando información como la descripción del sistema, su propósito previsto, el estado de conformidad y los datos de contacto del proveedor.

Este registro promueve la transparencia y les permite a las autoridades y al público monitorear qué sistemas de alto riesgo están en uso. También facilita la supervisión del mercado y la aplicación de medidas correctivas si es necesario.

3. Supervisión Humana.

Para los sistemas de IA de alto riesgo, el acta requiere medidas apropiadas de supervisión humana (Art. 14). Esto significa que estos sistemas deben diseñarse y desarrollarse de manera que permitan una supervisión efectiva por parte de personas naturales durante su uso. La supervisión humana es clave para detectar y mitigar riesgos, asegurar que los sistemas funcionen según lo previsto y que sus decisiones sean éticas y responsables.

4. Sistema de Gestión de Riesgos.

El acta exige que los proveedores de sistemas de IA de alto riesgo establezcan y documenten un sistema de gestión de riesgos (Art. 9). Este sistema debe:

- Identificar y analizar los riesgos conocidos y predecibles asociados con cada uso previsto del sistema a lo largo de su ciclo de vida,
- Estimar y evaluar los riesgos que pueden surgir, considerando la severidad del impacto y la probabilidad de ocurrencia,
- Adoptar medidas de mitigación adecuadas.

El sistema requerido es una herramienta importante para que los proveedores asuman la responsabilidad proactiva de abordar los riesgos potenciales de sus sistemas, en lugar de simplemente reaccionar a los problemas después del hecho.

Sin embargo, a pesar de las medidas previstas en el acta, determinar la responsabilidad por los resultados de los sistemas de IA sigue siendo un desafío complejo. Algunas de las dificultades incluyen:

- La IA a menudo implica cadenas de suministro y desarrollo complejas, con múltiples actores (proveedores de datos, desarrolladores de modelos, integradores de sistemas, usuarios finales, etc.) que pueden compartir la responsabilidad,
- Los sistemas de aprendizaje automático pueden exhibir comportamientos emergentes o inesperados que no fueron previstos o deseados por sus diseñadores²⁰²,
- Puede ser difícil entender o explicar cómo un sistema de IA llegó a una decisión particular, especialmente con técnicas de "caja negra".

Estos desafíos requieren un enfoque matizado de la responsabilidad que tenga en cuenta los diferentes roles y contribuciones de los actores involucrados, así como la naturaleza a veces opaca de la toma de decisiones de la IA.

J. Fomento de la Normalización y la Interoperabilidad

La normalización e interoperabilidad desempeñan un papel crucial en el desarrollo y despliegue de sistemas de Inteligencia Artificial (IA) confiables y seguros. Estos principios facilitan la compatibilidad, el intercambio de datos y la cooperación entre diferentes sistemas de IA, así como con otros productos y servicios digitales. En el contexto del Acta de IA de la Unión Europea, el fomento de la normalización y la interoperabilidad se considera esencial para garantizar un alto nivel de protección de los derechos fundamentales, la salud y la seguridad, al tiempo que se promueve la innovación y se fortalece el mercado único digital europeo

El acta de IA de la UE reconoce la importancia de estos principios y establece disposiciones específicas para promover el desarrollo y uso de normas armonizadas, especificaciones comunes y formatos de datos interoperables en el ámbito de la IA. Estas medidas tienen como objetivo facilitar el cumplimiento de los requisitos del acta, mejorar la transparencia y la trazabilidad de los sistemas de IA y fomentar la cooperación entre los distintos actores del ecosistema de IA.

La normalización se refiere al proceso de desarrollar y establecer requisitos, especificaciones, directrices o características técnicas que pueden ser utilizadas consistentemente para asegurar que los productos, servicios y procesos sean adecuados para su propósito²⁰³. En el

²⁰² Schaeffer, Miranda, y Koyejo, "Are Emergent Abilities," 2023.

²⁰³ Organización Internacional de Normalización. "Normas." Accedido el 26 de junio de 2024. <https://www.iso.org/standards.html>

contexto del Acta de IA, la normalización tiene como objetivo promover la seguridad, fiabilidad y consistencia de los sistemas de IA, así como facilitar su evaluación de conformidad con los requisitos establecidos.

El acta contiene varias disposiciones clave sobre normalización, especialmente en los Artículos 40 y 41. El Artículo 40 aborda las normas armonizadas y los artefactos de normalización, señalando que los sistemas de IA de alto riesgo o los modelos de IA de propósito general que cumplen con las normas armonizadas relevantes se presumirán conformes con ciertos requisitos del Acta. El Artículo 41, por su parte, permite a la comisión adoptar especificaciones comunes cuando las normas armonizadas no existan o sean insuficientes. Estas especificaciones comunes proporcionan medios para cumplir con los requisitos del acta.

El proceso de desarrollo de normas armonizadas y especificaciones comunes involucra a los organismos europeos de normalización, en consulta con partes interesadas y expertos relevantes. La comisión emite solicitudes de normalización a estos organismos, especificando los requisitos que deben cumplir las normas. Una vez adoptadas, las referencias a las normas armonizadas se publican en el Diario Oficial de la Unión Europea. El Artículo 42 establece que el cumplimiento de las normas armonizadas o especificaciones comunes confiere una presunción de conformidad con los requisitos correspondientes del acta, lo cual facilita la evaluación de conformidad de los sistemas de IA²⁰⁴.

La interoperabilidad, por otra parte, se refiere a la capacidad de diferentes sistemas, dispositivos o aplicaciones para conectarse e intercambiar información de manera efectiva y consistente²⁰⁵. En el ámbito de la IA, la interoperabilidad es crucial para permitir la comunicación y cooperación entre sistemas de IA, así como su integración con otros productos y servicios

²⁰⁴ **Artículo 42 - Presunción de conformidad con ciertos requisitos** 1. Los sistemas de IA de alto riesgo que hayan sido entrenados y probados con datos que reflejen el entorno geográfico, comportamental, contextual o funcional específico en el que se pretende que sean utilizados se presumirá que cumplen con los requisitos pertinentes establecidos en el artículo 10(4). 2. Los sistemas de IA de alto riesgo que hayan sido certificados o para los cuales se haya emitido una declaración de conformidad bajo un esquema de ciberseguridad conforme al Reglamento (UE) 2019/881 y cuyas referencias hayan sido publicadas en el Diario Oficial de la Unión Europea se presumirá que cumplen con los requisitos de ciberseguridad establecidos en el artículo 15 de este Reglamento, en la medida en que el certificado de ciberseguridad o la declaración de conformidad o partes de ellos cubran dichos requisitos.

²⁰⁵ Adrian Ostermann y Niklas Jooß, "Interoperability: definition, evaluation and application," FfE Munich, 16 de noviembre de 2022, <https://www.ffe.de/en/veroeffentlichungen/Interoperability-definition-evaluation-and-application>

digitales. Esto promueve la eficiencia, la innovación y la creación de ecosistemas digitales más cohesivos.

El Acta de IA establece varios requisitos de interoperabilidad para los sistemas de IA de alto riesgo. El Anexo IV²⁰⁶, que detalla el contenido de la documentación técnica, exige información sobre cómo el sistema de IA interactúa o puede ser utilizado para interactuar con hardware o software, incluidos otros sistemas de IA. Además, el Artículo 10 (3) aborda la interoperabilidad de los formatos de datos, requiriendo que los conjuntos de datos de entrenamiento, validación y prueba tengan las propiedades estadísticas apropiadas y sean relevantes para el propósito previsto del sistema.

El desarrollo de normas armonizadas y especificaciones comunes también contribuye a la interoperabilidad al proporcionar requisitos técnicos consistentes para la conectividad, el intercambio de datos y la compatibilidad entre sistemas de IA. Esto facilita la integración de

²⁰⁶ El Anexo IV del AI Act establece que la documentación técnica para los sistemas de inteligencia artificial (IA) debe ser exhaustiva y detallada, cubriendo múltiples aspectos esenciales para asegurar el cumplimiento de los estándares y regulaciones pertinentes. En primer lugar, se requiere una descripción general del sistema de IA, que debe incluir el propósito previsto del sistema, el nombre del proveedor y la versión del sistema, reflejando su relación con versiones anteriores. Además, debe detallarse cómo el sistema interactúa con otros hardware o software, incluidas otras IA, si procede, y las versiones relevantes de software o firmware junto con cualquier requisito relacionado con las actualizaciones. Esta sección también debe incluir una descripción de las formas en que el sistema se comercializa o se pone en servicio, las especificaciones del hardware en el que se ejecutará el sistema, imágenes o ilustraciones de los productos que incorporan el sistema de IA, y una descripción básica de la interfaz de usuario y las instrucciones de uso para el desplegador. En segundo lugar, la documentación técnica debe proporcionar una descripción detallada de los elementos del sistema de IA y del proceso de desarrollo. Esto incluye los métodos y pasos utilizados en el desarrollo del sistema, especialmente el uso de sistemas preentrenados o herramientas de terceros, así como las especificaciones de diseño del sistema, que abarcan la lógica general de la IA, las principales decisiones de diseño, las suposiciones realizadas y los parámetros optimizados. También es fundamental describir la arquitectura del sistema, los recursos computacionales utilizados, y los requisitos de datos, incluyendo las metodologías de entrenamiento y los conjuntos de datos empleados. Asimismo, se debe evaluar las medidas de supervisión humana necesarias y proporcionar una descripción de los cambios predeterminados en el sistema y su rendimiento, los procedimientos de validación y pruebas, y las medidas de ciberseguridad implementadas. Finalmente, la documentación debe incluir información detallada sobre la monitorización, el funcionamiento y el control del sistema de IA. Esto abarca las capacidades y limitaciones del rendimiento del sistema, los posibles resultados no deseados y las fuentes de riesgos para la salud, la seguridad y los derechos fundamentales, así como las medidas de supervisión humana necesarias. También se debe detallar la idoneidad de las métricas de rendimiento utilizadas, el sistema de gestión de riesgos aplicado durante todo el ciclo de vida del sistema, y los cambios relevantes realizados por el proveedor. Adicionalmente, se debe listar las normas armonizadas aplicadas y, en su defecto, describir las soluciones adoptadas para cumplir con los requisitos del reglamento, acompañadas de una copia de la declaración de conformidad de la UE. Finalmente, es necesario proporcionar una descripción detallada del sistema para evaluar el rendimiento del sistema de IA en la fase postcomercialización, incluyendo el plan de monitorización correspondiente.

sistemas de diferentes proveedores y promueve el desarrollo de soluciones de IA más versátiles y adaptables.

Los organismos europeos de normalización desempeñan un papel crucial en el desarrollo de normas armonizadas y especificaciones comunes para los sistemas de IA. Los tres organismos principales son el Comité Europeo de Normalización (CEN), el Comité Europeo de Normalización Electrotécnica (CENELEC) y el Instituto Europeo de Normas de Telecomunicaciones (ETSI). Estos organismos trabajan en estrecha colaboración con la Comisión Europea, los Estados miembros y las partes interesadas para desarrollar y promover la adopción de normas en el ámbito de la IA²⁰⁷.

El CEN y el CENELEC se centran en el desarrollo de normas en una amplia gama de sectores, incluyendo la IA aplicada a productos y servicios industriales, de consumo y de salud. El ETSI, por su parte, se especializa en normas para tecnologías de la información y la comunicación, incluyendo aspectos de la IA relacionados con las telecomunicaciones y la interoperabilidad de datos²⁰⁸.

Estos organismos de normalización colaboran estrechamente con expertos de la industria, la academia y la sociedad civil para garantizar que las normas desarrolladas sean técnicamente sólidas, respondan a las necesidades del mercado y tengan en cuenta consideraciones éticas y sociales. También participan en iniciativas internacionales de normalización para promover la alineación global de las normas de IA.

La normalización e interoperabilidad tienen un impacto significativo en el desarrollo y la adopción de sistemas de IA confiables y seguros. Al proporcionar requisitos técnicos consistentes y promover la compatibilidad entre sistemas, estas medidas contribuyen a la innovación, la competitividad y el crecimiento del mercado único digital europeo. Las normas armonizadas y las especificaciones comunes facilitan el cumplimiento de los requisitos regulatorios y simplifican el

²⁰⁷ CEN-CENELEC. "CEN-CENELEC Response to the EC White Paper on AI." Versión 2020-06. Accedido el 26 de junio de 2024. Recuperado de: https://www.cencenelec.eu/media/CEN-CENELEC/Areas%20of%20Work/Position%20Paper/cen-clc_ai_fg_white-paper-response_final-version_june-2020.pdf

²⁰⁸ ETSI. "Artificial Intelligence." Accedido el 26 de junio de 2024. <https://www.etsi.org/technologies/artificial-intelligence>

proceso de evaluación de conformidad, reduciendo las barreras para las empresas, especialmente las pymes¹⁵.

Sin embargo, el desarrollo e implementación de normas efectivas también presenta desafíos. El rápido ritmo de la innovación tecnológica en el campo de la IA puede dificultar el mantenimiento de normas actualizadas y relevantes. Además, lograr un consenso entre las diversas partes interesadas sobre los requisitos técnicos puede ser un proceso complejo y que requiere mucho tiempo. Es necesario encontrar un equilibrio adecuado entre la normalización y la flexibilidad para no obstaculizar la innovación.

2.3.3.- Análisis del AI Act en relación con la Implementación de Sistemas de IA en la Administración de Justicia

A. Clasificación de los Sistemas de IA Judiciales como de Alto Riesgo.

El Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial, comúnmente conocido como el AI Act, aborda, de manera específica, la implementación de sistemas de IA en el ámbito de la administración de justicia. En su Anexo III, punto 8(a), el AI Act clasifica expresamente como sistemas de IA de alto riesgo aquellos "*destinados a ser utilizados por una autoridad judicial o en su nombre para asistir a las autoridades judiciales en la investigación e interpretación de los hechos y el derecho y en la aplicación de la ley a un conjunto concreto de hechos*". Esta categorización evidencia el reconocimiento por parte del legislador europeo de los potenciales impactos significativos que el despliegue de sistemas de IA puede tener en este dominio sensible, especialmente en lo relativo a los derechos fundamentales, el Estado de Derecho y los principios democráticos (Considerando 48)²⁰⁹.

²⁰⁹ *La magnitud del impacto adverso causado por el sistema de IA en los derechos fundamentales protegidos por la Carta es de particular relevancia al clasificar un sistema de IA como de alto riesgo. Estos derechos incluyen el derecho a la dignidad humana, el respeto a la vida privada y familiar, la protección de los datos personales, la libertad de expresión e información, la libertad de reunión y asociación, el derecho a la no discriminación, el derecho a la educación, la protección del consumidor, los derechos de los trabajadores, los derechos de las personas con discapacidad, la igualdad de género, los derechos de propiedad intelectual, el derecho a un recurso efectivo y a un juicio justo, el derecho a la defensa y la presunción de inocencia, y el derecho a una buena administración. Además de esos derechos, es importante destacar el hecho de que los niños tienen derechos específicos consagrados en el Artículo 24 de la Carta y en la Convención de las Naciones Unidas sobre los Derechos del Niño, desarrollados en mayor medida en el Comentario General N° 25 de la CDN con respecto al entorno digital, ambos de los cuales*

La clasificación de los sistemas de IA judiciales como de alto riesgo se fundamenta en la naturaleza crítica de las funciones que desempeñan y en la magnitud de las consecuencias que podrían derivarse de su uso inadecuado o defectuoso. Como señala el Considerando 59, "*las acciones de las autoridades policiales que implican determinados usos de sistemas de IA se caracterizan por un grado significativo de desequilibrio de poder y pueden dar lugar a una vigilancia, detención o privación de libertad de una persona física, así como a otros efectos adversos sobre los derechos fundamentales*"²¹⁰. Esta consideración es igualmente aplicable al contexto judicial, donde la asistencia en la investigación e interpretación de hechos y derecho, así como en la aplicación de la ley a casos concretos, son tareas neurálgicas que inciden, directamente, en la tutela judicial efectiva de los justiciables. Errores o sesgos en estas actividades podrían conllevar vulneraciones a derechos fundamentales como la igualdad ante la ley (artículo 20 de la Carta de los Derechos Fundamentales de la Unión Europea), el derecho a la tutela judicial efectiva y a un juez imparcial (artículo 47 de la Carta), la presunción de inocencia y los derechos de la defensa (artículo 48 de la Carta) o el derecho a la libertad y a la seguridad (artículo 6 de la Carta).

Además, el empleo de sistemas de IA en la toma de decisiones judiciales plantea desafíos significativos en términos de transparencia, explicabilidad y rendición de cuentas. Como apunta el Considerando 61, "*para permitir la confianza en la actuación de las autoridades judiciales y garantizar la rendición de cuentas, es importante que los sistemas de IA utilizados en este contexto*

requieren considerar las vulnerabilidades de los niños y proporcionar la protección y el cuidado necesarios para su bienestar. El derecho fundamental a un alto nivel de protección ambiental consagrado en la Carta e implementado en las políticas de la Unión también debe considerarse al evaluar la gravedad del daño que un sistema de IA puede causar, incluyendo en relación con la salud y la seguridad de las personas.

²¹⁰ (59) *Dada su función y responsabilidad, las acciones de las autoridades de aplicación de la ley que involucren ciertos usos de sistemas de IA se caracterizan por un significativo grado de desequilibrio de poder y pueden conducir a la vigilancia, arresto o privación de la libertad de una persona natural, así como a otros impactos adversos en los derechos fundamentales garantizados en la Carta. En particular, si el sistema de IA no se entrena con datos de alta calidad, no cumple con los requisitos adecuados en términos de su rendimiento, precisión o robustez, o no está diseñado y probado adecuadamente antes de ser comercializado o puesto en servicio, puede señalarse a las personas de manera discriminatoria o incorrecta o injusta. Además, el ejercicio de importantes derechos fundamentales procesales, como el derecho a un recurso efectivo y a un juicio justo, así como el derecho de defensa y la presunción de inocencia, podría verse obstaculizado, en particular, cuando dichos sistemas de IA no son suficientemente transparentes, explicables y documentados. Por lo tanto, es apropiado clasificar como de alto riesgo, en la medida en que su uso esté permitido bajo la legislación pertinente de la Unión y nacional, a una serie de sistemas de IA destinados a ser utilizados en el contexto de la aplicación de la ley, donde la precisión, fiabilidad y transparencia son particularmente importantes para evitar impactos adversos, mantener la confianza pública y asegurar la responsabilidad y la reparación efectiva.*

*sean suficientemente transparentes, explicables y documentados*²¹¹. La opacidad de ciertos algoritmos y la dificultad para comprender su lógica interna podrían obstaculizar el ejercicio del derecho a un recurso efectivo y a un juicio justo, al dificultar que los afectados cuestionen las resoluciones basadas en IA (artículo 47 de la Carta). Asimismo, existe el riesgo de que se genere una excesiva deferencia hacia las recomendaciones algorítmicas, erosionando la independencia judicial (artículo 19 del Tratado de la Unión Europea) y la confianza pública en la justicia.

Por otra parte, la clasificación de alto riesgo busca prevenir la amplificación o perpetuación de sesgos discriminatorios históricos a través de los sistemas de IA. Dado que los algoritmos se entrena con datos que pueden reflejar prejuicios sociales preexistentes, existe el peligro de que se reproduzcan patrones de discriminación en función de la raza, el sexo, el origen étnico, la orientación sexual u otras características protegidas (Considerando 59). Esto podría dar lugar a decisiones judiciales injustas y socavar la igualdad ante la ley (artículo 20 de la Carta) y la prohibición de toda discriminación (artículo 21 de la Carta).

La categorización de los sistemas de IA judiciales como de alto riesgo implica que estarán sujetos a requisitos estrictos en materia de gestión de riesgos (artículo 9), gobierno de datos (artículo 10), documentación técnica (artículo 11), transparencia (artículo 13), supervisión humana (artículo 14), exactitud (artículo 15), robustez (artículo 15) y ciberseguridad (artículo 15). Estos requerimientos, detallados en la Sección 2 del Capítulo III del AI Act, tienen por objeto garantizar que el desarrollo y uso de dichos sistemas se ajuste a estándares rigurosos de calidad, seguridad y

²¹¹ (61) *Ciertos sistemas de IA destinados a la administración de justicia y a los procesos democráticos deben clasificarse como de alto riesgo, considerando su impacto potencialmente significativo en la democracia, el estado de derecho, las libertades individuales, así como el derecho a un recurso efectivo y a un juicio justo. En particular, para abordar los riesgos de posibles sesgos, errores y opacidad, es apropiado calificar como de alto riesgo a los sistemas de IA destinados a ser utilizados por una autoridad judicial o en su nombre para asistir a las autoridades judiciales en la investigación e interpretación de los hechos y el derecho, y en la aplicación del derecho a un conjunto concreto de hechos. Los sistemas de IA destinados a ser utilizados por órganos de resolución alternativa de disputas para esos fines también deben considerarse de alto riesgo cuando los resultados de los procedimientos de resolución alternativa de disputas produzcan efectos legales para las partes. El uso de herramientas de IA puede apoyar el poder de decisión de los jueces o la independencia judicial, pero no debe reemplazarlo: la toma de decisiones final debe seguir siendo una actividad impulsada por humanos. Sin embargo, la clasificación de los sistemas de IA como de alto riesgo no debe extenderse a los sistemas de IA destinados a actividades administrativas puramente auxiliares que no afecten la administración real de justicia en casos individuales, como la anonimización o pseudonimización de decisiones judiciales, documentos o datos, la comunicación entre el personal y las tareas administrativas.*

respeto de los derechos fundamentales. Solo así se podrá aprovechar el potencial de la IA para mejorar la eficiencia y calidad de la justicia, al tiempo que se mitigan sus riesgos inherentes.

B. Requisitos Específicos para Sistemas de IA de Alto Riesgo en el Ámbito Judicial.

El AI Act establece una serie de requisitos específicos que deberán cumplir los proveedores y usuarios de sistemas de IA de alto riesgo en el ámbito judicial, a fin de abordar los riesgos particulares que plantea su implementación en este contexto sensible. Estos requerimientos, recogidos principalmente en la Sección 2 del Capítulo III del Reglamento, abarcan aspectos como la gestión de riesgos, el gobierno de datos, la documentación técnica, la transparencia, supervisión humana, exactitud y ciberseguridad.

En primer lugar, el artículo 9 exige que los proveedores establezcan, implementen y documenten un sistema de gestión de riesgos que permita identificar, analizar, evaluar y mitigar los riesgos conocidos y razonablemente previsibles para la salud, la seguridad y los derechos fundamentales que puedan derivarse del uso de los sistemas de IA judiciales. Este proceso deberá llevarse a cabo de manera continua e iterativa a lo largo de todo el ciclo de vida del sistema (artículo 9.2) e incluir medidas apropiadas y específicas de gestión de riesgos, dando debida consideración a los efectos e interacciones resultantes de la aplicación combinada de los requisitos establecidos en la Sección 2 (artículo 9.4).

Asimismo, el artículo 10 establece exigencias rigurosas en materia de gobierno de datos. Los conjuntos de datos utilizados para el entrenamiento, la validación y prueba de los sistemas de IA judiciales deberán ser "*pertinentes, representativos, exentos de errores y completos*" (artículo 10.3), teniendo debidamente en cuenta las características específicas del contexto judicial en el que se prevé utilizar el sistema (artículo 10.4). Los proveedores deberán aplicar procedimientos apropiados de gestión de datos, incluyendo técnicas de examen, limpieza, agregación y retención de datos (artículo 10.2(f)), así como medidas para detectar y mitigar posibles sesgos que puedan afectar a la salud, la seguridad o los derechos fundamentales o dar lugar a discriminación prohibida por el Derecho de la Unión (artículo 10.2(g)).

Otro requisito clave es la elaboración y el mantenimiento de una documentación técnica detallada que demuestre el cumplimiento del sistema de IA con los requerimientos del AI Act.

Según el artículo 11 y el Anexo IV, esta documentación deberá incluir, entre otros elementos, una descripción general del sistema, sus especificaciones de diseño, los procedimientos de verificación y validación aplicados, así como información sobre su desempeño previsto en relación con su finalidad. Además, deberá mantenerse actualizada a lo largo de todo el ciclo de vida del sistema (artículo 11.1) y ponerse a disposición de las autoridades nacionales competentes.

La transparencia, como ya se ha explicado de manera exhaustiva, es otro aspecto fundamental abordado por el AI Act. El artículo 13.1 exige que los sistemas de IA judiciales estén diseñados y desarrollados de manera que su funcionamiento sea "*suficientemente transparente para permitir a los usuarios interpretarlos y utilizarlos adecuadamente*". Para ello, deberán ir acompañados de instrucciones de uso que incluyan información concisa, completa, correcta y clara sobre sus características, capacidades y limitaciones, las circunstancias que pueden dar lugar a riesgos, las medidas de supervisión humana implementadas y, cuando proceda, especificaciones sobre los datos de entrada (artículo 13.3).

Precisamente, la supervisión humana es otro requisito esencial para asegurar que las decisiones asistidas por IA estén sujetas al control último de un juez o funcionario judicial competente. El artículo 14 dispone que los sistemas de IA de alto riesgo se diseñarán y desarrollarán de manera que puedan ser efectivamente supervisados por personas físicas durante el periodo en que se utilicen. Las medidas de supervisión deberán ser apropiadas para prevenir o minimizar los riesgos, garantizar que el sistema esté sujeto a limitaciones operativas integradas que no puedan ser anuladas, sea receptivo al operador humano y les permita a las personas a las que se ha asignado la supervisión interpretar correctamente la salida del sistema, detectar errores o disfunciones y decidir no utilizarlo o anular sus resultados en una situación concreta (artículo 14.4). Además, en el caso de los sistemas enumerados en el Anexo III, punto 1(a), relativos a la identificación biométrica remota, se requerirá que la correspondencia arrojada por el algoritmo sea verificada y confirmada por al menos dos personas físicas (artículo 14.5).

El AI Act también exige que los sistemas de IA judiciales logren un nivel apropiado de exactitud, robustez y ciberseguridad, en consonancia con el estado reconocido de la técnica (artículo 15.1). Los niveles de exactitud deberán declararse en las instrucciones de uso (artículo 15.3) y los sistemas tendrán que ser resilientes en relación con errores, fallos o incoherencias

(artículo 15.4) y frente a intentos de alteración o manipulación malintencionada por parte de terceros no autorizados (artículo 15.5).

Asimismo, los proveedores deberán establecer un sistema de monitoreo poscomercialización que recabe y analice activamente datos sobre el desempeño de los sistemas de IA judiciales a lo largo de su ciclo de vida, a fin de garantizar el cumplimiento continuo de los requisitos del AI Act (artículo 72.2). En caso de identificar cualquier riesgo o incumplimiento serio, deberán informar inmediatamente a las autoridades de vigilancia del mercado y adoptar las acciones correctivas necesarias (artículo 20.1).

El AI Act establece un marco normativo exhaustivo y coherente que busca garantizar que el desarrollo y uso de sistemas de IA en el ámbito judicial se ajuste a los más altos estándares éticos y de protección de los derechos fundamentales. Los requisitos específicos impuestos a los sistemas de alto riesgo, que abarcan aspectos como la gestión de riesgos, el gobierno de datos, la transparencia, la supervisión humana y la exactitud, tienen por objeto mitigar los riesgos inherentes a la aplicación de estas tecnologías en un dominio tan sensible y salvaguardar los principios esenciales del Estado de Derecho. Solo así se podrá aprovechar el potencial de la IA para mejorar la calidad y eficiencia de la administración de justicia, al tiempo que se preserva la independencia judicial y la confianza ciudadana en las instituciones judiciales.

C. Garantías Procesales en el AI Act.

El AI Act establece un conjunto de garantías procesales específicas para los sistemas de IA de alto riesgo, con el objetivo de salvaguardar los derechos fundamentales y garantizar la confianza en la aplicación de estas tecnologías en ámbitos sensibles como la administración de justicia. Estas garantías se articulan en torno a tres ejes principales: la transparencia y explicabilidad, la supervisión humana y el derecho a la revisión humana de las decisiones.

C.1. Transparencia y Explicabilidad (arts. 13, 52)

El artículo 13(1) del AI Act establece que los sistemas de alto riesgo deben diseñarse y desarrollarse de manera que su funcionamiento sea suficientemente transparente para permitir que los usuarios interpreten la producción del sistema y la utilicen adecuadamente. Esta disposición

busca garantizar que los operadores jurídicos que empleen estos sistemas, como jueces, fiscales o abogados, puedan comprender su funcionamiento y fundamentar adecuadamente sus decisiones.

Para ello, los sistemas deben ir acompañados de instrucciones de uso que incluyan información concisa, completa, correcta y clara (art. 13(2)). Esta información debe abarcar las características, capacidades y limitaciones de rendimiento del sistema, incluyendo:

- Su propósito previsto,
- Su nivel de precisión y métricas de exactitud, robustez y ciberseguridad (art. 13(3)(b)(ii)),
- Cualquier circunstancia conocida o previsible que pueda dar lugar a riesgos para la salud, la seguridad o los derechos fundamentales (art. 13(3)(b)(iii)),
- Especificaciones sobre los datos de entrada u otra información pertinente sobre los conjuntos de datos utilizados (art. 13(3)(b)(vi)),
- Información para permitir a los usuarios interpretar la producción del sistema de IA (art. 13(3)(b)(vii)).

Esta información es crucial para que los operadores jurídicos puedan valorar adecuadamente los resultados arrojados por el sistema de IA y utilizarlos de manera apropiada en sus decisiones. Deben poder discernir en qué circunstancias el sistema puede ser fiable y en cuáles sus limitaciones o posibles sesgos desaconsejan su uso o exigen precauciones adicionales.

Además de estas obligaciones generales de transparencia, el artículo 50 establece requisitos específicos de información para los sistemas de IA utilizados por autoridades judiciales u otras autoridades públicas. En estos casos, se debe informar a las personas naturales afectadas de que están interactuando con un sistema de IA, a menos de que esto sea obvio desde el punto de vista de una persona razonablemente bien informada, observadora y cauta, teniendo en cuenta las circunstancias y el contexto de uso.

C.2. Supervisión Humana (art. 14)

Otra garantía procesal clave establecida en el AI Act es la exigencia de supervisión humana para los sistemas de IA de alto riesgo. El artículo 14(1) dispone que estos sistemas deben diseñarse y desarrollarse de manera que puedan ser efectivamente supervisados por personas físicas durante el período en que se utilicen.

Esta supervisión tiene como objetivo prevenir o minimizar los riesgos para la salud, la seguridad o los derechos fundamentales que puedan surgir cuando se utilice el sistema de IA (art. 14(2)). Se trata de asegurar que las decisiones no se deleguen por completo en los sistemas automatizados, sino que exista siempre un control y una supervisión por parte de operadores humanos.

Las medidas de supervisión deben ser proporcionales a los riesgos, el nivel de autonomía y el contexto de uso del sistema (art. 14(3)). Pueden incluir tanto medidas integradas en el propio diseño del sistema por el proveedor, como medidas a implementar por el usuario (art. 14(3)(a) y (b)). En todo caso, deben permitir a los supervisores humanos:

- Comprender plenamente las capacidades y limitaciones del sistema de IA y supervisar adecuadamente su funcionamiento (art. 14(4)(a)),
- Permanecer atentos ante posibles tendencias de confianza automática o excesiva en la producción del sistema (sesgo de automatización) (art. 14(4)(b)),
- Ser capaces de interpretar correctamente la producción del sistema, con las herramientas de interpretación disponibles (art. 14(4)(c)),
- Decidir, en cualquier situación particular, no utilizar el sistema de IA o ignorar, anular o revertir su producción (art. 14(4)(d)),
- Intervenir en el funcionamiento del sistema o interrumpirlo mediante un botón de "parada" u otro procedimiento similar (art. 14(4)(e)).

Estas previsiones buscan garantizar que los operadores humanos mantengan en todo momento el control último sobre las decisiones, pudiendo apartarse del criterio sugerido por el sistema de IA cuando lo consideren necesario. En el ámbito judicial, esto es fundamental para preservar la independencia judicial y evitar una excesiva deferencia hacia las recomendaciones algorítmicas.

En el caso de los sistemas biométricos utilizados por autoridades judiciales, se refuerza esta garantía exigiendo que ninguna acción o decisión se base en la identificación resultante del sistema a menos que haya sido verificada y confirmada por separado por al menos dos personas físicas (art. 14(5)).

C.3.- Derecho a la Revisión Humana de las Decisiones (art. 86)

El artículo 86 del AI Act establece el derecho a la revisión humana de las decisiones individuales tomadas con la asistencia de sistemas de IA de alto riesgo. Esta disposición constituye una garantía procesal fundamental para salvaguardar los derechos de las personas afectadas por decisiones automatizadas y asegurar una supervisión humana adecuada.

Según el artículo 86(1), **cualquier persona afectada por una decisión basada en un sistema de IA de alto riesgo enumerado en el Anexo III, que produzca efectos jurídicos o le afecte significativamente de modo similar, tendrá derecho a obtener del usuario una explicación clara y significativa del papel del sistema de IA en el procedimiento decisorio y de los principales elementos de la decisión adoptada.** Este derecho se activa cuando la persona considera que la decisión tiene un impacto adverso en su salud, seguridad o derechos fundamentales. Se trata de un umbral amplio, que no se limita a efectos jurídicos directos, sino que abarca también afectaciones significativas similares, permitiendo cubrir un espectro diverso de situaciones.

El derecho a la revisión humana opera como un mecanismo de control *ex post*, complementario a las garantías de transparencia y supervisión humana *ex ante*. Mientras que estas últimas buscan asegurar que las decisiones automatizadas sean tomadas de manera adecuada y con suficiente intervención humana desde el principio, el derecho a la revisión permite cuestionar decisiones ya adoptadas cuando la persona afectada considera que son injustas o incorrectas.

Este derecho adquiere especial relevancia en el contexto de la administración de justicia, donde las decisiones pueden tener un impacto directo y significativo en los derechos e intereses de las personas. Permite a las partes de un proceso obtener un escrutinio adicional de las decisiones que les perjudiquen, más allá del control que hayan podido ejercer previamente a través de las obligaciones de transparencia e información.

Además, el artículo 86(1) especifica que la revisión debe centrarse, tanto en el papel que ha jugado el sistema de IA en la toma de la decisión, como en los principales elementos de la decisión en sí misma. Esto implica un examen en profundidad del funcionamiento del sistema de

IA y de su influencia en el resultado, pero también una valoración global de la decisión, incluyendo los demás factores y las pruebas considerados.

Es importante destacar que el derecho a la revisión humana no implica necesariamente un derecho a obtener una decisión diferente. Su finalidad principal es garantizar un control adecuado del proceso decisorio y permitirle a la persona afectada exponer sus argumentos y cuestionar los elementos que considere incorrectos o injustos. La autoridad decisoria mantiene su discrecionalidad para confirmar o modificar la decisión inicial, pero debe hacerlo de manera motivada, abordando las alegaciones planteadas en la revisión.

Otro aspecto relevante es la relación entre el derecho a la revisión humana y las posibles vías de recurso judicial contra las decisiones basadas en IA. Aunque el artículo 86 no aborda directamente esta cuestión, cabe entender que la revisión opera como una garantía adicional a las vías de impugnación ordinarias; no como un requisito previo o sustitutivo de las mismas.

Por otro lado, el artículo 86(2) establece algunas limitaciones al derecho a la revisión humana, indicando que no se aplicará en los usos de sistemas de IA para los que la legislación de la Unión o nacional prevea su exclusión o restricción. Esta previsión busca salvaguardar aquellos casos en los que, por razones imperiosas de interés público u otras consideraciones legítimas, se permita un mayor grado de automatización de las decisiones. No obstante, dado el carácter fundamental del derecho a la revisión como garantía procesal, cabe esperar que estas excepciones sean interpretadas de manera restrictiva y proporcionada y que estén justificadas por razones sólidas y objetivas.

El derecho a la revisión humana de las decisiones basadas en IA se posiciona como un eje central para generar confianza en el uso de estas tecnologías en el ámbito judicial, asegurando que su aplicación se realiza de manera compatible con los principios de equidad, transparencia y respeto a los derechos de defensa. Su efectividad dependerá, en gran medida, de una adecuada implementación práctica por parte de los Estados miembros y las autoridades competentes, así como de una interpretación apropiada de sus interacciones con otras normas procesales aplicables.

2.4.- Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales y su Entorno

La Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales y su Entorno (en adelante, "la Carta") fue adoptada por la Comisión Europea para la Eficiencia de la Justicia (CEPEJ) en su 31^a Reunión Plenaria los días 3 y 4 de diciembre de 2018 en Estrasburgo²¹². La CEPEJ es un órgano del Consejo de Europa, la organización internacional que tiene como fin promover la democracia y proteger los derechos humanos y el Estado de Derecho en Europa.

Este instrumento del Consejo de Europa, cristalizó los principios rectores y valores fundamentales que, posteriormente, serían pormenorizados y dotados de fuerza vinculante por el legislador comunitario. La Carta Ética funcionó como un verdadero preámbulo normativo, estableciendo el sustrato ético-jurídico sobre el cual se erigiría la futura regulación. Sus cinco principios cardinales - respeto de los derechos fundamentales, no discriminación, calidad y seguridad, transparencia e imparcialidad, y control por el usuario - prefiguraron los ejes axiológicos que vertebran el EU AI Act. Este último, en su afán de proveer un marco regulatorio exhaustivo, viene a concretar y operativizar dichos principios mediante disposiciones específicas, requisitos técnicos y mecanismos de gobernanza.

Es menester destacar que la Carta Ética, al ser concebida en 2018, refleja, inexorablemente, el estado del arte de la tecnología de inteligencia artificial de aquel momento. Las consideraciones sobre las capacidades y limitaciones de los sistemas de IA plasmadas en el documento pueden haber quedado, en cierta medida, superadas por los vertiginosos avances acaecidos en el lustro subsiguiente. Verbigracia, las apreciaciones sobre la inviabilidad de ciertos usos de la IA en el ámbito judicial podrían requerir una revisión a la luz de los progresos en áreas como el procesamiento del lenguaje natural y el aprendizaje profundo.

²¹² European Commission for the Efficiency of Justice (CEPEJ). *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment*. Strasbourg: Council of Europe, 2018. Recuperado: <https://www.europarl.europa.eu/cmsdata/196205/COUNCIL%20OF%20EUROPE%20European%20Ethical%20Charter%20on%20the%20use%20of%20AI%20in%20judicial%20systems.pdf>

En particular, las aseveraciones relativas a la imposibilidad de los sistemas de IA para reproducir o modelar el razonamiento jurídico (párrafos 74-83) merecen ser reexaminadas. Como se expuso en acápitres anteriores, los desarrollos recientes en modelos de lenguaje de gran escala y técnicas de aprendizaje por refuerzo han expandido, significativamente, las capacidades de los sistemas de IA para comprender y generar texto jurídico complejo, así como para realizar tareas de razonamiento legal más sofisticadas.

2.4.1.- Naturaleza y Alcance Jurídico

Desde un punto de vista formal, la Carta se configura como un instrumento de *soft law* o derecho blando en el ámbito del Consejo de Europa. El *soft law* se define como aquel conjunto de instrumentos jurídicos que, si bien carecen de efectos vinculantes stricto sensu, están destinados a producir determinados efectos jurídicos y a influir en la conducta de los Estados, organizaciones internacionales y particulares²¹³. A diferencia de los tratados internacionales o las normas de derecho derivado de la Unión Europea, que son vinculantes para sus destinatarios, los instrumentos de *soft law* no tienen fuerza jurídica obligatoria directa. La persuasión que ejerce se basa en su autoridad moral e intelectual, además de su papel en la formulación de usos, costumbres y buenas prácticas²¹⁴.

No obstante, el *soft law* desempeña un papel relevante en la construcción progresiva del Derecho Internacional y europeo. Como señala Alonso García, el *soft law* anuncia o prepara en muchas ocasiones el *hard law*, permitiendo asentar unos principios que, posteriormente, cristalizarán en normas jurídicas vinculantes²¹⁵. Asimismo, el *soft law* posee una innegable fuerza persuasiva y un efecto didáctico, al orientar el comportamiento de los Estados y demás operadores jurídicos en una determinada dirección.

En este contexto, la Carta se erige como un documento de principios adoptado por un órgano especializado del Consejo de Europa con el fin de proporcionar un marco ético común para

²¹³ Federico Torrealba Navas, *Principios del Derecho Privado* (San José, Costa Rica: IJ Editores, Librería y Editorial Juricentro, abril 2021), p. 74.

²¹⁴ Ibid.

²¹⁵ Ricardo Alonso García, "El Soft Law Comunitario," *Revista de Administración Pública* no. 154 (2001): 74, cap. IV, "El soft law en cuanto avance del hard law". Recuperado de: <https://www.cepc.gob.es/sites/default/files/2021-12/243532001154063.pdf>

el desarrollo y uso de la inteligencia artificial en los sistemas judiciales de los Estados miembros de la organización. Aunque la Carta no es jurídicamente vinculante *per se*, sí proyecta una clara voluntad de la CEPEJ de sentar las bases axiológicas que deben guiar la implantación de estas tecnologías disruptivas en un ámbito tan sensible como la administración de justicia.

Según se desprende de su propio texto, la Carta está dirigida a los responsables públicos y privados encargados del diseño y despliegue de instrumentos y servicios de inteligencia artificial que impliquen el procesamiento de decisiones y datos judiciales, a los responsables públicos a cargo del marco legislativo y reglamentario, así como a las autoridades encargadas del desarrollo, auditoría o uso de tales instrumentos y servicio.

Por tanto, la Carta tiene como principales destinatarios a los poderes públicos de los Estados miembros del Consejo de Europa involucrados en la gobernanza de la IA en los sistemas judiciales (poder legislativo, ejecutivo y judicial), pero, también, interpela al sector privado en tanto que desarrollador y proveedor de soluciones de IA para la justicia. Nos encontramos, así, ante un instrumento omnicomprensivo que busca involucrar a todos los actores relevantes en este ámbito.

Aunque la Carta no tenga carácter obligatorio, su adopción por la CEPEJ, como organismo intergubernamental especializado, sí permite atribuirle ciertos efectos jurídicos. En primer lugar, la Carta despliega un efecto interpretativo en relación con otros instrumentos normativos aplicables en este campo, tanto a nivel del Consejo de Europa (en especial, el Convenio Europeo de Derechos Humanos), como a nivel de la Unión Europea (señaladamente, el Reglamento General de Protección de Datos y el Reglamento sobre la IA o EU AI Act). Los principios consagrados en la Carta pueden servir, así, como criterio hermenéutico a la hora de interpretar y aplicar estas normas en lo que respecta al uso de la IA en el contexto judicial.

En segundo lugar, la Carta puede actuar como parámetro de control de la actividad de los Estados y de los operadores privados en esta materia. Si bien el incumplimiento de las disposiciones de la Carta no puede ser directamente sancionado, sí cabe concebir que genere una cierta presión política y un reproche desde el punto de vista del respeto a los estándares europeos plasmados en esta. El apartado sobre la "Aplicación de la Carta" prevé precisamente que las autoridades independientes mencionadas puedan evaluar periódicamente el nivel de cumplimiento

de los principios por parte de los distintos actores y proponer mejoras para su adaptación. La Carta advierte incluso de que los principios "*deben ser objeto de una aplicación, un seguimiento y una evaluación regulares*" por los actores públicos y privados, quienes "*deberán explicar, en su caso, las razones por las que no se aplican o se aplican parcialmente, acompañadas de un plan de acción para introducir las medidas necesarias*"²¹⁶. Se configura, así, un incipiente mecanismo de rendición de cuentas basado en el cumplimiento o explicación (*comply or explain*²¹⁷).

En suma, aunque la Carta no sea una norma vinculante en sentido estricto, resulta evidente que está llamada a irradiar importantes efectos jurídicos y a convertirse en uno de los principales marcos de referencia ético en el ámbito europeo para el desarrollo y uso responsable de la inteligencia artificial en la administración de justicia. Su dimensión axiológica y su vocación de implementación práctica la dotan de una especial *auctoritas* que infundirá – como ya lo hizo con el EU AI Act - sin duda los desarrollos legislativos y las aplicaciones concretas de la IA en este sector en los años venideros.

2.4.2.- Análisis de los Principios Rectores de la Carta

El eje vertebrador de la Carta lo conforman cinco principios basilares que deben regir la concepción, el desarrollo y la utilización de los sistemas de inteligencia artificial en el ámbito judicial

➤ Principio de Respeto de los Derechos Fundamentales

En primer lugar, la Carta establece el respeto de los derechos fundamentales como premisa inexcusable. Según este principio, el diseño y la aplicación de herramientas y servicios de IA debe ser plenamente compatible con los derechos fundamentales garantizados en el Convenio Europeo de Derechos Humanos, la Carta de los Derechos Fundamentales de la UE y el *corpus iuris* en materia de protección de datos personales²¹⁸.

²¹⁶ European Commission for the Efficiency of Justice (CEPEJ). *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment*, p. 5.

²¹⁷ Thomson Reuters. "Comply or Explain Approach." *Practical Law Glossary*, Resource ID 8-107-5967, 2024. [https://uk.practicallaw.thomsonreuters.com/8-107-5967?transitionType=Default&contextData=\(sc.Default\)&firstPage=true](https://uk.practicallaw.thomsonreuters.com/8-107-5967?transitionType=Default&contextData=(sc.Default)&firstPage=true)

²¹⁸ European Commission for the Efficiency of Justice (CEPEJ). *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment*, p. 8.

Esto implica adoptar un enfoque "*ético y respetuoso con los derechos humanos desde la fase de diseño*" (*ethical-by-design* y *human rights-by-design*). Es decir, los derechos y garantías fundamentales han de ser tomados en consideración e incorporados desde las etapas iniciales de concepción de los sistemas, no como un añadido o corrección *a posteriori*.

Este principio exige que los actores involucrados en el desarrollo de aplicaciones de IA para la justicia lleven a cabo una evaluación rigurosa y continua del impacto potencial de estas tecnologías sobre los derechos humanos, con especial atención a los derechos de defensa, el derecho a un proceso equitativo, la igualdad de armas procesales, el derecho a un tribunal independiente e imparcial y la presunción de inocencia, entre otros.

Asimismo, el principio de respeto de los derechos fundamentales demanda que la implementación de sistemas de IA en la justicia incorpore desde el diseño las debidas garantías frente a los riesgos de erosión de la privacidad y la protección de datos personales. Deben aplicarse, escrupulosamente, los principios y las salvaguardas del Reglamento General de Protección de Datos de la UE, como la minimización de datos, la privacidad por defecto y desde el diseño, evaluaciones de impacto, etc.

En este sentido, la Carta sugiere anonimizar o pseudonimizar los datos judiciales utilizados en el desarrollo de modelos predictivos, para evitar la reidentificación de los justiciables. También apunta la necesidad de una base jurídica adecuada para el tratamiento de categorías especiales de datos, como los datos biométricos o sobre condenas penales²¹⁹.

➤ Principio de No Discriminación

El segundo principio de la Carta exige prevenir, específicamente, el desarrollo o la intensificación de cualquier discriminación entre individuos o grupos de individuos. Dado que los sistemas de IA tienen la capacidad de revelar discriminaciones existentes mediante la agrupación o clasificación de datos relativos a personas o grupos de personas, los actores públicos y privados

²¹⁹ Ibid, p. 26.

deben asegurarse de que los métodos no reproduzcan o agraven tales discriminaciones y no conduzcan a análisis deterministas o utilizaciones deterministas²²⁰.

Se requiere una especial cautela, tanto en la fase de desarrollo, como de despliegue, cuando el tratamiento se base, directa o indirectamente, en datos "sensibles" como el origen racial o étnico, las opiniones políticas, las convicciones religiosas o filosóficas, la afiliación sindical, los datos genéticos o biométricos, los datos relativos a la salud o a la orientación sexual. No obstante, el uso del aprendizaje automático y de análisis científicos multidisciplinares para luchar contra tales discriminaciones debe fomentarse.

Este principio entraña con la prohibición de discriminación consagrada en el artículo 14 del CEDH y en el Protocolo nº 12 a este. Tiene también una estrecha relación con las disposiciones del Reglamento General de Protección de Datos (RGPD) que prohíben las decisiones individuales automatizadas, incluida la elaboración de perfiles, que produzcan efectos jurídicos o afecten, significativamente, a las personas basándose únicamente en categorías especiales de datos, salvo que se apliquen las garantías adecuadas (artículo 22 del RGPD²²¹).

➤ Principio de Calidad y Seguridad

De acuerdo con el tercer principio de la Carta, en lo que respecta al tratamiento de resoluciones y datos judiciales, deben utilizarse fuentes certificadas y datos intangibles con modelos concebidos de manera multidisciplinar, en un entorno tecnológico seguro [^14]. Ello implica que los diseñadores de los modelos de aprendizaje automático puedan basarse, ampliamente, en los conocimientos especializados de los profesionales del sistema judicial (jueces,

²²⁰ Ibid, p. 10.

²²¹ **Artículo 22. Decisiones individuales automatizadas, incluida la elaboración de perfiles.** *Todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar. El apartado 1 no se aplicará si la decisión: a) es necesaria para la celebración o la ejecución de un contrato entre el interesado y un responsable del tratamiento; b) está autorizada por el Derecho de la Unión o de los Estados miembros que se aplique al responsable del tratamiento y que establezca asimismo medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado; o c) se basa en el consentimiento explícito del interesado. En los casos a que se refiere el apartado 2, letras a) y c), el responsable del tratamiento adoptará las medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado, como mínimo el derecho a obtener intervención humana por parte del responsable, a expresar su punto de vista y a impugnar la decisión. Las decisiones a que se refiere el apartado 2 no se basarán en las categorías especiales de datos personales contempladas en el artículo 9, apartado 1, salvo que se aplique el artículo 9, apartado 2, letra a) o g), y se hayan tomado medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado.*

fiscales, abogados, etc.) y de los investigadores en el ámbito del Derecho y las ciencias sociales. La constitución de equipos de proyecto mixtos que trabajen en ciclos cortos permite capitalizar este enfoque multidisciplinar.

Además, los datos judiciales que se introduzcan en un programa de aprendizaje automático deben proceder de fuentes certificadas y no deben modificarse hasta su utilización efectiva por el mecanismo de aprendizaje. Todo el proceso debe ser trazable para garantizar que no se ha producido ninguna modificación que altere el contenido o sentido de la decisión procesada. Los modelos y algoritmos creados también deben poder almacenarse y ejecutarse en entornos seguros, a fin de garantizar la integridad e intangibilidad del sistema²²².

➤ **Principio de Transparencia, Imparcialidad y Equidad**

El cuarto principio proclama que se debe lograr un equilibrio entre la propiedad intelectual de ciertos métodos de procesamiento y la necesidad de transparencia (posibilidad de acceso al proceso de diseño), imparcialidad (ausencia de sesgo), equidad e integridad intelectual (prioridad a los intereses de la justicia) cuando se utilizan herramientas que pueden tener repercusiones jurídicas o afectar, significativamente, la vida de las personas²²³.

La Carta plantea como primera opción la transparencia técnica total (por ejemplo, el código fuente abierto y documentación), si bien reconoce que a veces puede estar restringida por derechos de propiedad industrial e intelectual. También admite la posibilidad de que el responsable explique, en un lenguaje claro y comprensible, la naturaleza de los servicios ofrecidos, las herramientas desarrolladas, sus resultados y los riesgos de error. Asimismo, sugiere que autoridades u organismos expertos independientes certifiquen y auditen los métodos de tratamiento o proporcionen asesoramiento previo.

Se trata de un principio clave para generar confianza en el uso de la IA en el ámbito judicial, que conecta con los principios de buena administración, el derecho a la tutela judicial efectiva y las garantías procesales reconocidas por el artículo 6 del CEDH (en concreto, el derecho a un tribunal independiente e imparcial y a un proceso equitativo). La Carta apunta así a la necesidad

²²² Ibid, p. 11.

²²³ Ibid, p. 12.

de conciliar la protección de los derechos de propiedad industrial de los desarrolladores de sistemas de IA con las exigencias de transparencia, control y rendición de cuentas inherentes a la administración de justicia.

➤ **Principio de Control por parte del Usuario**

Finalmente, el quinto principio afirma que debe excluirse un enfoque prescriptivo y garantizar que los usuarios son actores informados y controlan sus elecciones. Los profesionales en el sistema de justicia deberían, en cualquier momento, poder revisar las decisiones judiciales y los datos utilizados para producir un resultado y no estar necesariamente obligados a seguirlo a la luz de las características específicas de ese caso en particular²²⁴.

El usuario debe ser informado, en un lenguaje claro y comprensible, de si las soluciones ofrecidas por los instrumentos de IA son vinculantes o no, de las diferentes opciones posibles y de su derecho a asesoramiento jurídico y a acceder a un tribunal.

Además, cuando se implante un sistema de información basado en IA, deben establecerse programas de formación en informática para los usuarios y fomentarse los debates con los profesionales del mundo jurídico. Se trata, en definitiva, de preservar la autonomía del usuario y evitar que la utilización de sistemas de IA pueda traducirse en una merma del poder decisorio del juez o del derecho a la tutela judicial efectiva de los justiciables.

2.4.3.- Apéndices de la Carta Ética: Análisis y Contribuciones Esenciales

Los apéndices de la Carta Ética Europea constituyen un corpus documental de insoslayable valor, complementando y profundizando los principios rectores establecidos en el cuerpo principal del instrumento. Estos anexos proporcionan un marco analítico y operativo exhaustivo para la implementación de sistemas de inteligencia artificial (IA) en el ámbito judicial.

²²⁴ Ibid, p. 13.

1. Apéndice I: Estudio en Profundidad sobre el Uso de la IA en los Sistemas Judiciales²²⁵

El Apéndice I de la Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales y su Entorno constituye un estudio pormenorizado y multidimensional que aborda, con rigor científico y perspectiva jurídica, la compleja interrelación entre las tecnologías de inteligencia artificial y la administración de justicia en el contexto europeo. Este análisis jurídico-tecnológico comienza delineando el *status quo* (para el 2018) de la implementación de algoritmos de IA en los sistemas judiciales de los Estados miembros del Consejo de Europa, revelando una dicotomía significativa entre el interés creciente por estas tecnologías disruptivas y su todavía incipiente aplicación práctica en el ámbito jurisdiccional. El estudio subraya, con especial énfasis, la preponderancia de iniciativas provenientes del sector privado, lo cual suscita interrogantes de calado sobre la necesidad imperiosa de una mayor intervención y regulación por parte de los poderes públicos en este campo neurálgico para el Estado de Derecho.

En su análisis de las políticas de datos abiertos relativas a las decisiones judiciales, el apéndice aborda una cuestión de capital importancia para el desarrollo de sistemas de IA en el ámbito judicial: la disponibilidad y accesibilidad de los datos jurisprudenciales. Este examen no se limita a una mera descripción del *status quo*, sino que ahonda en las implicaciones jurídicas y éticas de la apertura de estos datos, poniendo de relieve los potenciales conflictos con el derecho fundamental a la protección de datos personales. En este contexto, el estudio hace especial hincapié en la necesidad perentoria de implementar medidas robustas de anonimización y pseudonimización, como salvaguarda inexcusable para la protección de los derechos de los justiciables.

El apéndice procede a realizar una disección minuciosa de las características operativas de la IA aplicada a las decisiones judiciales, centrándose, particularmente, en las técnicas de aprendizaje automático. Este análisis técnico-jurídico reviste especial relevancia, pues desvela la naturaleza esencialmente estadística y correlacional de estos sistemas, desmitificando la noción de que la IA pudiese, en su estado actual al 2018, replicar o sustituir el razonamiento jurídico humano. Esta constatación sirve de fundamento para una reflexión más amplia sobre las limitaciones

²²⁵ Ibid, p. 16-62.

intrínsecas de la IA en la modelización del razonamiento jurídico, poniendo de manifiesto la complejidad inherente a la labor hermenéutica y argumentativa que caracteriza la praxis jurídica.

El estudio no se circunscribe a un análisis teórico, sino que explora con detenimiento las aplicaciones potenciales de la IA en diversos ámbitos de la administración de justicia, con especial atención a la justicia civil, comercial y administrativa. En este contexto, se examinan propuestas innovadoras como la elaboración de escalas jurisprudenciales mediante técnicas de IA o la implementación de sistemas de resolución alternativa de conflictos en línea asistidos por inteligencia artificial. No obstante, el apéndice adopta una postura de prudencia epistemológica, subrayando la necesidad imperiosa de mantener la supervisión humana y de implementar salvaguardas efectivas para prevenir una excesiva automatización de la función jurisdiccional.

En lo concerniente a la justicia penal, el estudio aborda con particular cautela y profundidad analítica los desafíos específicos que plantea la utilización de sistemas de IA en este ámbito especialmente sensible. Se examinan críticamente las propuestas de utilización de algoritmos para la prevención de delitos y la evaluación de riesgos de reincidencia, poniendo de relieve los peligros latentes de perpetuación y amplificación de sesgos discriminatorios. El apéndice advierte, con fundamentada preocupación, sobre el riesgo de que estas tecnologías puedan socavar principios fundamentales del Derecho penal, como la presunción de inocencia o la individualización de la pena.

La protección de datos personales emerge como un leitmotiv a lo largo del estudio, reconociendo su papel crucial en la intersección entre IA y justicia. El apéndice realiza un análisis pormenorizado de las implicaciones del uso de la IA en el tratamiento de datos personales en el ámbito judicial, subrayando la necesidad inexcusable de observar escrupulosamente los principios rectores establecidos en el Reglamento General de Protección de Datos y en el Convenio 108+ del Consejo de Europa. Se hace especial hincapié en la necesidad de implementar medidas técnicas y organizativas que garanticen la integridad, confidencialidad y disponibilidad de los datos, así como el respeto a los derechos de los interesados.

2. Apéndice II: ¿Qué usos de la IA Considerar en los sistemas Judiciales Europeos?²²⁶

El Apéndice II de la Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales y su Entorno proporciona una taxonomía de los diversos usos de la inteligencia artificial (IA) en los sistemas judiciales europeos, estableciendo una gradación en cuanto a la conveniencia de su implementación a la luz de los principios y valores consagrados en la Carta Ética.

Este apéndice, lejos de ofrecer una visión maniquea sobre la aplicación de la IA en el ámbito judicial, presenta un enfoque matizado que reconoce, tanto las potencialidades, como los riesgos inherentes a estas tecnologías disruptivas. En este sentido, el documento identifica cuatro categorías de usos, a saber:

- usos a fomentar,
- usos posibles que requieren precauciones metodológicas considerables,
- usos a considerar después de estudios científicos adicionales,
- usos a considerar con las reservas más extremas.

Esta clasificación no solo ofrece una hoja de ruta para los operadores jurídicos y los responsables de la formulación de políticas públicas, sino que, también, sirve como un marco analítico para evaluar la idoneidad y los potenciales impactos de las aplicaciones de IA en el ecosistema judicial:

Categoría	Descripción
Usos por fomentar	En esta categoría, el apéndice identifica aplicaciones de IA que no solo son compatibles con los principios éticos establecidos, sino que, además, ofrecen un potencial significativo para mejorar la administración de justicia. Destaca, en primer lugar, la utilización de técnicas de aprendizaje automático para la creación y mejora de motores de búsqueda jurisprudencial. Esta aplicación representa un avance sustancial en la

²²⁶ Ibid, p. 63-69.

	<p>capacidad de los operadores jurídicos para acceder y analizar el corpus jurisprudencial, facilitando una interpretación más holística y coherente del derecho.</p> <p>Asimismo, se promueve el desarrollo de herramientas de IA para facilitar el acceso al derecho, incluyendo la implementación de <i>chatbots</i> capaces de interactuar en lenguaje natural para guiar a los usuarios a través de las diversas fuentes de información jurídica. Este uso democratiza el acceso a la justicia, reduciendo las barreras de entrada para los ciudadanos legos en derecho.</p> <p>Otro uso fomentado es la creación de nuevas herramientas estratégicas basadas en la ciencia de datos y la IA para la gestión judicial. Estas aplicaciones permiten realizar evaluaciones cuantitativas y cualitativas del funcionamiento de los tribunales, así como proyecciones sobre recursos humanos y presupuestarios. No obstante, el apéndice subraya acertadamente la importancia de involucrar a los profesionales del derecho, especialmente a los jueces, en la implementación y análisis de estos instrumentos, a fin de salvaguardar la calidad de la justicia y el acceso a esta.</p>
Usos posibles que requieren precauciones metodológicas considerables	<p>Esta categoría aborda aplicaciones de IA que, si bien ofrecen beneficios potenciales, conllevan riesgos significativos que requieren una implementación cautelosa y metodológicamente rigurosa. Un ejemplo paradigmático es la asistencia en la elaboración de escalas en ciertos litigios civiles. El apéndice advierte acertadamente sobre la necesidad de identificar todos los factores causales, tanto explícitos, como implícitos, en las decisiones judiciales para evitar sesgos estadísticos que podrían perpetuar o exacerbar inequidades existentes.</p>

En el ámbito de la resolución alternativa de conflictos, se contempla el uso de IA como apoyo, pero se subraya la importancia de la transparencia en los sistemas de cálculo y la intervención de terceros capacitados. Esta precaución es crucial para mantener la integridad y la equidad de los procesos de resolución de conflictos.

En cuanto a la resolución de litigios en línea, el apéndice enfatiza la necesidad de informar claramente a los litigantes sobre la naturaleza del procesamiento de su disputa y garantizar la posibilidad de recurrir a un tribunal real en cumplimiento del artículo 6 del Convenio Europeo de Derechos Humanos. Esta salvaguarda es esencial para prevenir la erosión del derecho a un juicio justo en la era digital.

En el ámbito penal, se considera el uso de algoritmos en la investigación criminal para identificar la comisión de delitos, aunque se advierte sobre los riesgos de efectos performativos²²⁷ y la necesidad de un enfoque más amplio que

²²⁷ El concepto de "efecto performativo o de autorrealización" se refiere a la posibilidad de que un sistema, especialmente uno automatizado o basado en inteligencia artificial, comience a generar resultados que se autorrefuerzan o se repiten debido a la forma en que influye en las personas que proporcionan la información de entrada.

Para entenderlo mejor, pensemos en un sistema de escalas judiciales. Estas escalas son herramientas que los jueces pueden usar para tomar decisiones en casos legales. Si un sistema de IA se basa en decisiones previas que fueron tomadas utilizando estas escalas, puede empezar a producir resultados que son representativos solo de las decisiones anteriores, en lugar de considerar nuevos factores o contextos. En otras palabras, el sistema "aprende" a repetir patrones pasados sin introducir variaciones o ajustes basados en nuevas realidades o circunstancias.

Este ciclo puede hacer que el sistema se vuelva cerrado en sí mismo, produciendo resultados que perpetúan una forma específica de pensar o actuar, sin cuestionar si esos resultados siguen siendo adecuados o justos. Así, el sistema no evoluciona ni mejora, sino que simplemente repite el mismo tipo de salida una y otra vez, porque las entradas que recibe están cada vez más influenciadas por los resultados que el mismo sistema ha generado previamente. Esto puede llevar a una falta de diversidad en los resultados y, en el peor de los casos, a la perpetuación de sesgos o injusticias.

	<p>incluya a fiscales y académicos. Esta cautela refleja una comprensión sofisticada de las complejidades inherentes a la aplicación de la IA en el ámbito penal.</p>
Usos por considerar después de estudios científicos adicionales	<p>Esta categoría reconoce el potencial de ciertas aplicaciones de IA, pero subraya la necesidad de una base empírica más sólida antes de su implementación a gran escala. El perfilado de jueces es un ejemplo paradigmático: el apéndice sugiere que, en lugar de buscar sesgos personales, se ofrezca a los magistrados evaluaciones cuantitativas y cualitativas más detalladas de su actividad con fines puramente informativos. Esta aproximación refleja una comprensión matizada de la complejidad de la toma de decisiones judiciales y los riesgos de una simplificación excesiva.</p> <p>La anticipación de decisiones judiciales es otro uso que se ubica en esta categoría. El apéndice señala acertadamente las limitaciones de los enfoques puramente estadísticos y aboga por el desarrollo de sistemas híbridos que incorporen modelos matemáticos más sofisticados. Esta recomendación refleja una apreciación de la naturaleza multifacética del razonamiento jurídico y la insuficiencia de enfoques reduccionistas.</p>
Usos por considerar con las reservas más extremas	<p>En esta categoría, el apéndice identifica aplicaciones de IA que plantean riesgos significativos para los derechos fundamentales y la integridad del sistema judicial. El uso de algoritmos para perfilar individuos en el ámbito penal es objeto de una crítica particularmente incisiva. Citando experimentos controvertidos como COMPAS en Estados Unidos y HART en el Reino Unido, el documento advierte sobre el peligro de perpetuar sesgos discriminatorios y socavar principios fundamentales del derecho penal europeo, como la rehabilitación y la reintegración.</p>

Asimismo, se rechaza categóricamente la noción de una norma basada en la mera cantidad de decisiones previas. El apéndice argumenta, persuasivamente, que la acumulación de precedentes no puede suplantar o complementar la ley, y destaca los peligros de la cristalización de la jurisprudencia y los potenciales efectos negativos sobre la imparcialidad e independencia judicial. Esta posición refleja una comprensión profunda de la naturaleza dinámica y contextual del derecho, resistiendo la tentación de reducir la complejidad jurídica a meras estadísticas.

En última instancia, se puede concluir que el Apéndice II de la Carta Ética Europea sobre el Uso de la IA en los sistemas judiciales ofrece un marco de referencia invaluable para la implementación prudente y éticamente responsable de tecnologías de IA en el ámbito judicial. Su enfoque matizado y multidimensional reconoce, tanto el potencial transformador de estas tecnologías, como los riesgos inherentes a su adopción acrítica. Al establecer esta taxonomía cuatripartita, el apéndice no solo proporciona una guía práctica para los operadores jurídicos y los responsables de políticas públicas, sino que, también, sienta las bases para un diálogo continuo sobre la intersección entre la IA y el derecho, promoviendo así una modernización reflexiva y éticamente fundamentada de los sistemas judiciales en la era digital.

3. Apéndice III: Glosario²²⁸

El Apéndice III de la Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los sistemas judiciales y su entorno constituye un glosario técnico que proporciona definiciones precisas de los términos utilizados en el documento principal, con el objetivo de garantizar una comprensión unívoca y coherente de los conceptos fundamentales relacionados con la inteligencia artificial (IA) y su aplicación en el ámbito judicial. Este compendio terminológico resulta de vital importancia para la correcta interpretación y aplicación de los principios y las directrices

²²⁸ Ibid, p. 69-76.

establecidos en la Carta, toda vez que la materia abordada se caracteriza por su elevada complejidad técnica y la constante evolución de las tecnologías implicadas.

Entre las definiciones más relevantes que ofrece el glosario, cabe destacar la distinción entre la inteligencia artificial "fuerte" y "débil", siendo esta última la que se podía implementar en el 2018 en los sistemas judiciales. La IA "débil" se caracteriza por su alto rendimiento en ámbitos específicos de entrenamiento, sin llegar a emular la capacidad cognitiva humana en su totalidad. Esta precisión terminológica resulta crucial para delimitar el alcance y las expectativas respecto de las capacidades reales de los sistemas de IA en el contexto judicial.

Asimismo, el glosario aborda conceptos esenciales como el aprendizaje automático (machine learning), que constituye el núcleo de las aplicaciones de IA en el ámbito jurídico. Se define como un proceso mediante el cual se construye un modelo matemático a partir de datos, incorporando un gran número de variables no conocidas a priori. Esta definición técnica permite comprender el fundamento de las herramientas de IA utilizadas para el análisis jurisprudencial y la predicción de decisiones judiciales.

El documento también clarifica la noción de "justicia predictiva", concepto controvertido que se refiere al análisis de grandes volúmenes de decisiones judiciales mediante tecnologías de IA con el fin de realizar predicciones sobre el resultado de ciertos tipos de litigios especializados. El glosario advierte sobre las limitaciones y críticas a este enfoque, subrayando la complejidad inherente a la modelización matemática de fenómenos sociales.

En lo concerniente a la protección de datos personales, el glosario proporciona definiciones precisas de conceptos como anonimización y seudonimización, fundamentales para garantizar el respeto a la privacidad en el tratamiento de información judicial mediante sistemas de IA. Estas definiciones se alinean con el marco normativo europeo en materia de protección de datos, en particular con el Reglamento General de Protección de Datos (RGPD).

El Apéndice III también esclarece la distinción entre datos abiertos (*open source*) y la mera información pública unitaria disponible en sitios web, subrayando que los datos abiertos implican

la posibilidad de descarga y reutilización de bases de datos estructuradas, sujetas a licencias específicas. Esta distinción es crucial para comprender el alcance y las implicaciones de las políticas de transparencia judicial en el contexto de la implementación de sistemas de IA.

En síntesis, el glosario contenido en el Apéndice III de la Carta Ética Europea sobre el Uso de la IA en los Sistemas Judiciales se presenta un instrumento hermenéutico indispensable para la correcta comprensión y aplicación de los principios éticos y jurídicos que deben regir la implementación de tecnologías de IA en el ámbito de la administración de justicia.

4. Apéndice IV: Lista de Verificación para Integrar los Principios de la Carta en su Método de Procesamiento²²⁹

El Apéndice IV de la Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los sistemas judiciales y su entorno es un instrumento de autoevaluación de singular relevancia para los operadores jurídicos y desarrolladores de sistemas de inteligencia artificial en el ámbito judicial. Este apéndice, denominado *Lista de verificación para integrar los principios de la Carta en su método de procesamiento*, proporciona un mecanismo pragmático y sistemático para evaluar el grado de adherencia de las metodologías de procesamiento de datos a los principios rectores establecidos en la Carta.

La estructura del instrumento se fundamenta en una escala de autoevaluación que permite a los usuarios calibrar, de manera granular y objetiva, la conformidad de sus procesos con cada uno de los principios cardinales enunciados en el corpus principal del documento. El *modus operandi* propuesto consiste en la asignación de valoraciones en una escala gradual, donde el extremo izquierdo denota una integración plena del principio en cuestión, mientras que el extremo derecho indica una ausencia total de implementación.

Este enfoque metodológico facilita una apreciación matizada y diferenciada del nivel de cumplimiento, evitando, así, una dicotomía simplista entre conformidad y no conformidad. La culminación del proceso evaluativo se materializa en una sumatoria cuantitativa que, si bien no constituye una certificación formal, proporciona un indicador heurístico del grado global de

²²⁹ Ibid, p. 76-77.

alineación con los preceptos de la Carta. Es menester subrayar que este mecanismo de autoevaluación, lejos de ser un ejercicio meramente declarativo, se erige como un instrumento de reflexión crítica y mejora continua para los actores involucrados en el desarrollo y la aplicación de tecnologías de inteligencia artificial en el ecosistema judicial.

La lista de verificación aborda, de manera pormenorizada, los cinco principios medulares consagrados en la Carta, a saber: el respeto de los derechos fundamentales, la no discriminación, la calidad y seguridad, la transparencia, imparcialidad y equidad y el control por parte del usuario. Para cada uno de estos principios, se proporciona un espectro evaluativo que permite a los usuarios calibrar con precisión el grado de integración de dichos principios en sus metodologías de procesamiento.

Este enfoque holístico y multidimensional permite una evaluación comprehensiva que trasciende los aspectos meramente técnicos, abarcando consideraciones éticas, jurídicas y sociales de cardinal importancia en el contexto de la administración de justicia. La implementación de este instrumento de autoevaluación no solo fomenta una cultura de responsabilidad y diligencia entre los desarrolladores y usuarios de sistemas de inteligencia artificial en el ámbito judicial, sino que, también, promueve una convergencia progresiva hacia estándares éticos y operativos armonizados a nivel europeo.

2.5.- Reglamento General de Protección de Datos (GDPR)

2.5.1.- Naturaleza Jurídica y Ámbito de Aplicación

El Reglamento General de Protección de Datos (RGPD) de la Unión Europea, que entró en vigor el 25 de mayo de 2018²³⁰, constituye el marco normativo central en materia de privacidad y protección de datos personales en el contexto europeo. Su naturaleza de reglamento le confiere aplicabilidad directa en todos los Estados miembros, garantizando así un elevado grado de armonización y coherencia regulatoria en todo el territorio de la Unión.

²³⁰ 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), Diario Oficial de la Unión Europea L 119 (4 de mayo de 2016): 1-88. Recuperado de: <https://www.boe.es/doue/2016/119/L00001-00088.pdf>

El RGPD establece un conjunto comprehensivo de normas relativas al tratamiento de datos personales, fijando los principios rectores, derechos de los interesados, obligaciones de responsables y encargados, así como los mecanismos para asegurar su cumplimiento efectivo. Su ámbito de aplicación es deliberadamente amplio, cubriendo, tanto el sector público, como privado, y aplicándose al tratamiento total o parcialmente automatizado de datos personales, así como al tratamiento no automatizado de datos personales contenidos o destinados a ser incluidos en un fichero.

Si bien el RGPD no aborda de manera específica la cuestión de la inteligencia artificial, muchas de sus disposiciones resultan altamente relevantes y aplicables en este contexto. La definición expansiva de "*tratamiento*" del artículo 4(2), que incluye "*cualquier operación o conjunto de operaciones realizadas sobre datos personales o conjuntos de datos personales, ya sea por procedimientos automatizados o no, como la recogida, registro, organización, estructuración, conservación, adaptación o modificación, extracción, consulta, utilización, comunicación por transmisión, difusión o cualquier otra forma de habilitación de acceso, cotejo o interconexión, limitación, supresión o destrucción*", claramente engloba los diversos usos de datos personales en sistemas de IA, desde el entrenamiento de modelos hasta la toma de decisiones automatizadas.

En este sentido, el RGPD proporciona un sólido sustrato normativo sobre el cual se debe asentar cualquier regulación sectorial de la Unión en lo relativo al tema *sub examine*. Los principios, derechos y obligaciones del RGPD marcan límites infranqueables y fijan estándares mínimos que deben ser respetados por cualquier despliegue de IA que involucre datos personales, sin perjuicio de normas adicionales o más específicas (como las analizadas en acápite supra) que puedan establecerse atendiendo a los riesgos particulares de cada ámbito.

Por tanto, una adecuada comprensión de la interacción entre el RGPD y la regulación de la IA resulta fundamental para abordar los desafíos que plantea esta tecnología desde una perspectiva de derechos fundamentales y protección de los datos personales.

2.5.2.- Principios de Protección de Datos y su Aplicación a la IA

Los principios de protección de datos establecidos en el artículo 5 del RGPD constituyen los cimientos sobre los que se construye todo el edificio normativo del reglamento. Su aplicación en el contexto de la inteligencia artificial resulta crucial para garantizar un desarrollo y uso de esta tecnología respetuoso con los derechos fundamentales.

En primer lugar, el principio de licitud, lealtad y transparencia exige que los datos personales sean tratados de manera lícita, leal y transparente en relación con el interesado (art. 5.1.a) RGPD). La licitud implica que el tratamiento debe basarse en alguna de las bases jurídicas previstas en el artículo 6, aspecto que abordaremos en el siguiente apartado. La lealtad y transparencia, por su parte, se refieren a la necesidad de informar claramente a los interesados sobre el uso que se dará a sus datos y las consecuencias de dicho tratamiento.

En el caso de la IA, esto plantea el reto de explicar, de forma comprensible para el público general, el funcionamiento de algoritmos que pueden ser altamente complejos y opacos, especialmente aquellos basados en redes neuronales profundas. Implicando así que los órganos judiciales que empleen sistemas de IA en la Unión deberán hacer un esfuerzo pedagógico para comunicar, en un lenguaje accesible, qué datos se utilizan, para qué fines, cómo se toman las decisiones y qué implicaciones pueden tener para los justiciables.

Por su parte, el principio de limitación de la finalidad implica que los datos deben ser recogidos con fines determinados, explícitos y legítimos, y no ser tratados ulteriormente de manera incompatible con dichos fines (art. 5.1.b) RGPD). Este principio puede entrar en tensión con algunas prácticas comunes en el desarrollo de sistemas de IA, como el uso de grandes conjuntos de datos (*datasets*) recopilados, originalmente, para diversos fines y su reutilización para entrenar modelos predictivos.

En el ámbito judicial, resulta entonces esencial definir de antemano, y de forma precisa, las finalidades para las que se recaban y tratan los datos personales (por ejemplo, para analizar patrones de litigiosidad, predecir riesgos de reincidencia, asistir en la toma de decisiones sobre medidas cautelares, etc.), además de asegurar que cualquier tratamiento posterior sea compatible con esos fines o cuente con una nueva base legítima. El uso de datos judiciales para entrenar

modelos de IA con finalidades distintas (por ejemplo, comerciales) sería claramente contrario a este principio.

En tercer lugar, el principio de minimización de datos supone que los datos deben ser adecuados, pertinentes y limitados a lo necesario en relación con los fines para los que son tratados (art. 5.1.c) RGPD). Esto implica que los responsables deben hacer un juicio crítico sobre qué datos son realmente necesarios para alcanzar sus objetivos legítimos y abstenerse de recopilar o conservar datos excesivos o irrelevantes.

Aplicado a la IA judicial, este principio obligaría a hacer una cuidadosa selección de las fuentes de datos y variables por utilizar en los algoritmos, descartando aquellas que no aporten un valor añadido claro para la finalidad perseguida o que puedan introducir sesgos discriminatorios. Por ejemplo, para un sistema de evaluación de riesgo de reincidencia, podría cuestionarse la pertinencia de utilizar datos como la etnia, el código postal o los antecedentes familiares del acusado, frente a factores más directamente relacionados con la conducta individual.

El principio de exactitud exige que los datos sean exactos y, si fuera necesario, actualizados, debiendo adoptarse todas las medidas razonables para que se supriman o rectifiquen sin dilación los datos personales que sean inexactos con respecto a los fines para los que se tratan (art. 5.1.d) RGPD). La exactitud de los datos es crucial para el buen funcionamiento de los sistemas de IA, ya que los errores o imprecisiones en los datos de entrenamiento pueden traducirse en predicciones o decisiones equivocadas.

En el contexto judicial, esto implica la necesidad de mecanismos robustos de control de calidad y auditoría de los datos utilizados para entrenar y alimentar los algoritmos, así como procesos ágiles para corregir errores cuando sean detectados. Dada la sensibilidad de las decisiones judiciales y su impacto sobre los derechos de las personas, el estándar de exactitud exigible a los sistemas de IA en este ámbito debe ser particularmente elevado.

Finalmente, el principio de integridad y confidencialidad implica que los datos deben ser tratados de tal manera que se garantice una seguridad adecuada, incluida la protección contra el tratamiento no autorizado o ilícito y contra su pérdida, destrucción o daño accidental, mediante la

aplicación de medidas técnicas u organizativas apropiadas (art. 5.1.f) RGPD). Este principio entraña con las obligaciones de seguridad previstas en los artículos 32 a 34 del RGPD.

En el ámbito de la IA judicial, garantizar la integridad y confidencialidad de los datos resulta crucial, habida cuenta de la sensibilidad de la información tratada (datos penales, de salud, etc.) y de las graves consecuencias que podría tener su revelación o manipulación indebida. Los sistemas de IA deberán incorporar medidas de seguridad reforzadas, como el encriptado, técnicas de anonimización o seudonimización, control de accesos, monitorización de amenazas y planes de respuesta a incidentes y violaciones de seguridad.

Extrapolado a la IA en el sector judicial, este principio implica que no basta con proclamar que un sistema es conforme con el RGPD, sino que deben implementarse mecanismos efectivos de control interno, documentación, evaluación de impacto (art. 35 RGPD), auditoría y rendición de cuentas.

2.5.3.- Bases Jurídicas para el Tratamiento y Decisiones Automatizadas

El artículo 6 del RGPD establece un listado taxativo de bases jurídicas que pueden legitimar el tratamiento de datos personales. Todo tratamiento debe basarse en al menos una de estas causas de licitud, que funcionan como condiciones habilitantes. En el contexto de la aplicación de sistemas de inteligencia artificial en la administración de justicia, dos de estas bases jurídicas revisten especial relevancia: la misión de interés público o el ejercicio de poderes públicos y el cumplimiento de una obligación legal aplicable al responsable.

La base jurídica de la misión de interés público o el ejercicio de poderes públicos (art. 6.1.e) RGPD) es probablemente la más pertinente para el uso de IA por parte de los órganos judiciales. Esta base permite el tratamiento cuando es necesario para el cumplimiento de una misión realizada en interés público o en el ejercicio de poderes públicos conferidos al responsable. La administración de justicia es, sin duda, una misión de interés público esencial en un Estado de Derecho y los órganos judiciales tienen atribuidos por ley poderes públicos para el ejercicio de la función jurisdiccional. No obstante, la mera invocación del interés público no es suficiente para justificar cualquier tratamiento de datos personales mediante sistemas de IA.

El considerando 45 del RGPD aclara que el tratamiento debe tener una base en el Derecho de la Unión o de los Estados miembros, y que dicha base debe determinar la finalidad del tratamiento, los criterios para su licitud y especificar las normas precisas que rijan la delimitación de las funciones y competencias del responsable. Por tanto, sería necesario que la legislación procesal o las normas de organización judicial previeran expresamente la posibilidad de utilizar sistemas de IA para fines concretos (por ejemplo, para analizar jurisprudencia, asistir en la valoración de pruebas o predecir riesgos), estableciendo las garantías adecuadas.

Además, el tratamiento basado en el interés público debe respetar el principio de proporcionalidad. Como señala el considerando 45, debe existir una relación clara y directa entre el tratamiento y la finalidad perseguida, y el responsable debe evaluar si existen medios menos intrusivos para alcanzar dicha finalidad con igual eficacia. Así pues, el uso de IA en la justicia a partir de este marco normativo debería limitarse a aquellos casos en que resulte estrictamente necesario y proporcionado, descartándose cuando los objetivos puedan lograrse por medios menos invasivos de la privacidad.

En algunos casos, el tratamiento de datos personales por sistemas de IA judiciales podría basarse también en el cumplimiento de una obligación legal aplicable al responsable (art. 6.1.c) RGPD). Esta base legitima los tratamientos necesarios para cumplir una obligación legal impuesta por el Derecho de la Unión o de los Estados miembros, siempre que dicha norma determine las finalidades y los medios del tratamiento o los criterios para su determinación (considerando 45). Por ejemplo, si una ley procesal obligara a utilizar sistemas de IA para ciertas actuaciones (como la asignación automatizada de casos o la anonimización de resoluciones), el tratamiento encontraría su base en dicha norma.

Adicionalmente, cuando el uso de sistemas de IA implique la toma de decisiones individuales automatizadas, incluida la elaboración de perfiles, deberán observarse las garantías especiales previstas en el artículo 22 del RGPD. **Este precepto reconoce a los interesados el derecho a no ser objeto de decisiones basadas únicamente en el tratamiento automatizado que produzcan efectos jurídicos en ellos o les afecten significativamente de modo similar.**

Este derecho, configurado como una prohibición para los responsables, admite excepciones cuando la decisión es necesaria para celebrar o ejecutar un contrato, está autorizada por el Derecho

aplicable al responsable, o se basa en el consentimiento explícito del interesado (art. 22.2 RGPD). De estos supuestos, solo la autorización legal parece aplicable en el contexto judicial, pues difícilmente la relación entre el justiciable y el órgano judicial puede calificarse de contractual y el consentimiento del interesado no parece una base apropiada dada la disparidad de poder existente.

Así pues, para que un sistema de IA pueda tomar decisiones individuales automatizadas en el ámbito judicial, debe existir una norma con rango de ley que lo autorice y que establezca medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado (art. 22.2.b) RGPD). Estas medidas deben incluir, como mínimo, el derecho a obtener intervención humana por parte del responsable, a expresar el propio punto de vista y a impugnar la decisión (art. 22.3 RGPD).

Esto implica que las decisiones plenamente automatizadas, sin supervisión humana, deben ser la excepción en la administración de justicia. Los sistemas de IA podrán utilizarse como herramientas de apoyo a la toma de decisiones, pero, difícilmente, podrán sustituir por completo la valoración última del juez. Además, el considerando 71 añade otras garantías frente a las decisiones automatizadas, como la obligación del responsable de utilizar procedimientos matemáticos o estadísticos adecuados, aplicar medidas que corrijan inexactitudes y minimicen el riesgo de errores y asegurar los datos de modo que se tengan en cuenta los riesgos para los intereses y derechos del interesado y se impidan efectos discriminatorios. Cuando la decisión automatizada se base en categorías especiales de datos (origen étnico, opiniones políticas, afiliación sindical, religión, vida sexual, datos genéticos o biométricos, etc.), solo será lícita con el consentimiento explícito del interesado o por razones de interés público esencial (art. 22.4 RGPD).

Por otra parte, cuando los sistemas de IA se utilicen no para tomar decisiones automatizadas, sino para realizar evaluaciones o predicciones que sirvan de apoyo a la decisión humana (por ejemplo, para evaluar el riesgo de reincidencia o de incomparecencia de un investigado), también deberán respetarse ciertas cautelas. Aunque, en estos casos, no sería aplicable el régimen especial del artículo 22, sí lo serían los principios generales de protección de datos, como la transparencia, minimización de datos o la exactitud.

Así, los órganos judiciales que utilicen estos sistemas deberán informar claramente a los interesados de su existencia y de la lógica subyacente, además de permitirles acceder a los datos utilizados para realizar la evaluación. También deberán asegurarse de que los datos empleados son adecuados, pertinentes y limitados a lo necesario, y de que son exactos y están actualizados. Si la evaluación se basa en un perfil, el interesado tendrá derecho a conocerlo y a oponerse a él (art. 21.1 RGPD).

2.5.4.- Salvaguardas y Garantías del GDPR Aplicadas a la IA en el Ámbito Jurisdiccional

a.- Derechos de los Interesados y su Ejercicio Frente a Sistemas de IA

El RGPD reconoce a los interesados un catálogo de derechos en relación con el tratamiento de sus datos personales, incluyendo los derechos de acceso, rectificación, supresión ("derecho al olvido"), limitación del tratamiento, portabilidad y oposición (arts. 15-21 RGPD). Estos derechos son directamente aplicables frente a tratamientos realizados por sistemas de IA, aunque su ejercicio puede presentar desafíos prácticos.

Por ejemplo, el derecho de acceso implica que el interesado tendrá derecho a obtener, del responsable, confirmación de si se están tratando o no datos personales que le conciernen y, en tal caso, derecho de acceso a los datos y a cierta información, como los fines del tratamiento, las categorías de datos, los destinatarios, el plazo de conservación previsto y la existencia de decisiones automatizadas, incluida la elaboración de perfiles (art. 15.1 RGPD).

En el contexto de la IA judicial, esto podría abarcar el acceso no solo a los datos de entrada utilizados para entrenar o alimentar el sistema, sino, también, a los perfiles o inferencias generadas por el algoritmo sobre el interesado. Garantizar un acceso efectivo requerirá soluciones técnicas y organizativas por parte de los responsables.

El derecho de oposición, por su parte, otorga al interesado el derecho a oponerse en cualquier momento, por motivos relacionados con su situación particular, a que datos personales que le conciernen sean objeto de tratamiento basado en la misión de interés público o el interés legítimo del responsable, incluida la elaboración de perfiles (art. 21.1 RGPD). El responsable debe dejar de tratar los datos, salvo que acredite motivos legítimos imperiosos que prevalezcan sobre los intereses, derechos y las libertades del interesado.

b.- Obligaciones de los Responsables y Encargados

El RGPD impone diversas obligaciones a los responsables del tratamiento, definidos como las personas físicas o jurídicas, autoridades públicas, servicios u otros organismos que, solo, o junto con otros, determinan los fines y medios del tratamiento de datos personales (art. 4.7 RGPD²³¹). Muchas de estas obligaciones son especialmente relevantes en el contexto de sistemas de IA.

El responsable debe aplicar medidas técnicas y organizativas apropiadas a fin de garantizar y poder demostrar que el tratamiento es conforme con el reglamento, teniendo en cuenta la naturaleza, ámbito, contexto y fines del tratamiento, así como los riesgos para los derechos y las libertades de las personas (art. 24.1 RGPD). Esto exige un enfoque proactivo de gestión de riesgos a lo largo de todo el ciclo de vida de los sistemas de IA.

En particular, los responsables están obligados a llevar a cabo, antes del tratamiento, una evaluación del impacto que supondrá para los derechos y las libertades de los interesados (art. 35 RGPD). Esta evaluación de impacto relativa a la protección de datos (EIPD) es obligatoria cuando el tratamiento entraña un alto riesgo, lo que ocurrirá en muchos casos de aplicación de IA en el ámbito judicial, dada la sensibilidad de los datos tratados (datos penales, de salud, etc.) y el potencial impacto significativo sobre los derechos de los justiciables.

La EIPD debe incluir una descripción sistemática de las operaciones de tratamiento previstas y de los fines del tratamiento, una evaluación de la necesidad y proporcionalidad de las operaciones con respecto a su finalidad, una evaluación de los riesgos para los derechos y las libertades de los interesados y las medidas previstas para afrontar los riesgos y demostrar la conformidad con el RGPD (art. 35.7). En el caso de sistemas de IA, la EIPD deberá prestar especial atención a aspectos como la calidad y representatividad de los datos de entrenamiento, la transparencia y explicabilidad de los modelos, la equidad y no discriminación en los resultados y la robustez y ciberseguridad del sistema.

²³¹ “(...) 7) «**responsable del tratamiento**» o «**responsable**»: la persona física o jurídica, autoridad pública, servicio u otro organismo que, solo o junto con otros, determine los fines y medios del tratamiento; si el Derecho de la Unión o de los Estados miembros determina los fines y medios del tratamiento, el responsable del tratamiento o los criterios específicos para su nombramiento podrá establecerlos el Derecho de la Unión o de los Estados miembros”.

Si la EIPD muestra que el tratamiento entrañaría un alto riesgo si no se adoptan medidas para mitigarlo, el responsable deberá consultar a la autoridad de control antes de proceder (art. 36.1 RGPD). Esta consulta previa brinda una oportunidad para que la autoridad examine la conformidad del tratamiento y formule recomendaciones, especialmente cuando se trate de tecnologías innovadoras como la IA. En el ámbito judicial, este mecanismo de supervisión *ex ante* puede resultar especialmente valioso para prevenir usos desproporcionados o discriminatorios de los algoritmos.

Otra obligación clave de los responsables es la de protección de datos desde el diseño y por defecto (art. 25 RGPD). Esto implica que deben aplicarse, tanto en el momento de determinar los medios de tratamiento, como durante el propio tratamiento, medidas técnicas y organizativas apropiadas concebidas para aplicar los principios de protección de datos de forma efectiva. Los sistemas de IA deberán diseñarse desde el inicio con salvaguardas de privacidad integradas, en lugar de tratar de incorporarlas *a posteriori*. Por ejemplo, técnicas como la seudonimización pueden utilizarse para proteger la identidad de los justiciables en los *datasets* utilizados para entrenar los algoritmos. Por su parte, la minimización de datos de la que habla el artículo puede aplicarse seleccionando cuidadosamente las variables que realmente aportan valor al modelo y descartando datos irrelevantes o desproporcionados. Todas estas "salvaguardas por diseño" deben considerarse desde las primeras etapas del desarrollo de los sistemas de IA judicial.

Los responsables también deben adoptar medidas para garantizar que, por defecto, solo se traten los datos personales necesarios para cada fin específico (art. 25.2 RGPD). En el contexto de la IA, esto puede implicar ajustes en la recogida de datos (evitando recopilar más datos de los estrictamente necesarios), en la configuración de los parámetros de los algoritmos (por ejemplo, estableciendo umbrales más altos para la retención o el uso de datos sensibles) o en las interfaces de usuario (evitando mostrar por defecto información personal innecesaria).

Otras obligaciones relevantes de los responsables en relación con la IA son la de llevar un registro de las actividades de tratamiento bajo su responsabilidad (art. 30 RGPD), la de cooperar con las autoridades de supervisión (art. 31 RGPD), la de implementar medidas de seguridad técnicas y organizativas apropiadas para garantizar un nivel de seguridad adecuado al riesgo (art. 32 RGPD), la de notificar las violaciones de seguridad a la autoridad de control y, en ciertos casos, a los interesados (arts. 33-34 RGPD) y la de designar un delegado de protección de datos cuando una organización, ya sea pública o privada, lleva a cabo un tratamiento de datos personales que

implica un seguimiento habitual y sistemático a gran escala, o maneja categorías especiales de datos sensibles, como los relacionados con condenas e infracciones penales. Este delegado debe ser seleccionado por su conocimiento especializado en protección de datos y puede formar parte del personal de la organización o trabajar mediante un contrato externo. Además, sus datos de contacto deben ser accesibles públicamente y comunicados a la autoridad de control correspondiente. También, un grupo empresarial o varias autoridades públicas pueden nombrar un único delegado, siempre que sea fácilmente accesible para todos los establecimientos o entidades involucradas (art. 37 RGPD).

En particular, la obligación de registro de actividades de tratamiento (art. 30 RGPD) puede resultar especialmente útil en el contexto de la IA judicial, al fomentar la trazabilidad y la documentación de todo el ciclo de vida de los algoritmos. Este registro debería incluir información sobre los fines del tratamiento, las categorías de interesados y de datos personales tratados, los destinatarios de los datos, las transferencias internacionales, los plazos de conservación y una descripción general de las medidas técnicas y organizativas de seguridad.

Además de estas obligaciones generales, el RGPD también impone obligaciones específicas a los responsables que realicen tratamientos que entrañen decisiones automatizadas con efectos significativos para los interesados (art. 22 RGPD). Como ya se ha indicado, en estos casos deberán implementarse medidas adecuadas para salvaguardar los derechos de los interesados, incluyendo la intervención humana, la expresión del propio punto de vista y la impugnación de la decisión. Cuando estas decisiones se basen en categorías especiales de datos, solo podrán adoptarse con el consentimiento explícito del interesado o por razones importantes de interés público (art. 22.4 RGPD).

Finalmente, no podemos olvidar las obligaciones que el RGPD impone a los encargados del tratamiento, definidos como las personas físicas o jurídicas, autoridades públicas, servicios u otros organismos que traten datos personales por cuenta del responsable (art. 4.8 RGPD). En el contexto de la IA, será habitual que los responsables (por ejemplo, el Poder Judicial) recurran a proveedores externos especializados para el desarrollo, entrenamiento y mantenimiento de los sistemas algorítmicos. Estos proveedores actuarán como encargados y deberán cumplir las obligaciones previstas en el artículo 28 del RGPD.

Entre estas obligaciones, destaca la de tratar los datos únicamente siguiendo instrucciones documentadas del responsable, la de garantizar que las personas autorizadas para tratar datos

personales se hayan comprometido a respetar la confidencialidad, la de aplicar las medidas técnicas y organizativas apropiadas para garantizar un nivel de seguridad adecuado al riesgo, la de asistir al responsable en la atención al ejercicio de los derechos de los interesados y la de suprimir o devolver los datos al responsable una vez finalice la prestación de los servicios (art. 28.3 RGPD).

Además, cuando un encargado vaya a subcontratar parte del tratamiento, deberá informar previamente al responsable y obtener su autorización (art. 28.2 RGPD). Esta autorización podrá ser genérica, pero, en tal caso, el encargado informará al responsable de cualquier cambio, dándole la oportunidad de oponerse. En todo caso, el subencargado quedará obligado por las mismas condiciones y garantías que el encargado principal.

En el ámbito de la IA judicial, será importante que los contratos entre los tribunales (responsables) y los proveedores tecnológicos (encargados) reflejen claramente estas obligaciones y prevean mecanismos de supervisión y auditoría para asegurar su cumplimiento. Los tribunales deberán velar por que los algoritmos desarrollados por terceros respeten plenamente los principios y derechos del RGPD y porque cualquier transferencia ulterior de los datos (por ejemplo, a subencargados) se realice con las debidas garantías.

c.- Códigos de Conducta y Certificación

El RGPD, en sus artículos 40 a 43, introduce dos instrumentos de autorregulación y certificación que pueden desempeñar un papel relevante en el fomento de buenas prácticas y la demostración de conformidad en el uso de sistemas de inteligencia artificial: los códigos de conducta y los mecanismos de certificación.

Los códigos de conducta son conjuntos de normas y compromisos voluntarios que los responsables o encargados del tratamiento pueden suscribir para concretar y complementar la aplicación del RGPD en un determinado sector o actividad. Estos códigos deben ser elaborados por las asociaciones y otros organismos representativos de categorías de responsables o encargados, en consulta con las partes interesadas y teniendo en cuenta las especificidades de cada ámbito (art. 40.2 RGPD).

En el contexto de la IA judicial, los códigos de conducta podrían servir para establecer directrices más específicas sobre cómo aplicar los principios y las obligaciones del RGPD al desarrollo y uso de sistemas algorítmicos en la administración de justicia. Estos códigos podrían abordar cuestiones como la evaluación de la calidad y sesgo de los datos de entrenamiento, la

transparencia y explicabilidad de los modelos, la equidad y no discriminación en las decisiones, la implementación de salvaguardas adecuadas frente a las decisiones automatizadas, o la gestión de los riesgos de ciberseguridad.

La adhesión a un código de conducta aprobado puede ser utilizada por los responsables y encargados como un elemento para demostrar el cumplimiento de las obligaciones del RGPD (art. 24.3 RGPD). Además, los códigos pueden prever mecanismos específicos de supervisión de su cumplimiento, a cargo de un organismo acreditado (art. 41 RGPD), lo que refuerza su eficacia y credibilidad.

Los mecanismos de certificación, por su parte, son procedimientos por los cuales un organismo independiente acreditado evalúa y certifica que un determinado producto, servicio o proceso cumple ciertos requisitos o estándares en materia de protección de datos. Estos mecanismos, que pueden incluir sellos y marcas, tienen, como finalidad, permitir a los interesados evaluar rápidamente el nivel de protección de datos de los productos y servicios, y a los responsables y encargados demostrar su conformidad con el RGPD (art. 42.1 RGPD).

En el ámbito *sub examine*, la certificación podría aplicarse tanto a los propios sistemas algorítmicos (por ejemplo, para acreditar que cumplen ciertos estándares de transparencia, equidad o robustez) como a los procesos de desarrollo y gobernanza de estos sistemas (por ejemplo, para verificar que se han realizado evaluaciones de impacto adecuadas o que se han implementado medidas apropiadas de protección de datos desde el diseño).

La obtención de una certificación aprobada puede ser utilizada por los responsables y encargados como un elemento para demostrar el cumplimiento de las obligaciones del RGPD, en particular en lo que respecta a la aplicación de las medidas técnicas y organizativas apropiadas y a la realización de evaluaciones de impacto relativas a la protección de datos (arts. 24.3, 25.3 y 35.8 RGPD).

En nuestro contexto temático, sería recomendable que el desarrollo de estos instrumentos se realizase en estrecha colaboración entre las autoridades judiciales, las autoridades de protección de datos, los proveedores tecnológicos, la academia y la sociedad civil. De este modo, se podrían establecer estándares y buenas prácticas que, por un lado, tengan en cuenta las especificidades y garantías propias de la función jurisdiccional y, por otro, aprovechen el conocimiento técnico y la experiencia práctica de los diferentes actores implicados.

Es importante tener en cuenta que ni los códigos de conducta ni las certificaciones pueden sustituir o rebajar las obligaciones y responsabilidades establecidas por el RGPD. Su función es complementar y especificar la aplicación de las normas, nunca desplazarlas o debilitarlas. Los responsables y encargados que suscriban un código u obtengan una certificación seguirán siendo plenamente responsables del cumplimiento del RGPD y deberán estar en condiciones de demostrarlo (art. 42.4 RGPD).

Además, la eficacia de estos instrumentos dependerá, en gran medida, de la calidad de sus contenidos, de la independencia y rigor de los mecanismos de supervisión y control, y del nivel de adhesión y compromiso de los actores implicados. Un código de conducta o una certificación que se perciban como meros ejercicios cosméticos o de "lavado de cara" pueden tener el efecto contrario al deseado, socavando la confianza de los interesados y del público en general.

En definitiva, los códigos de conducta y los mecanismos de certificación previstos en el RGPD ofrecen una oportunidad para complementar y reforzar la aplicación de las normas de protección de datos en el desarrollo y uso de sistemas de IA en la administración de justicia. Bien diseñados y aplicados, estos instrumentos pueden aportar una mayor seguridad jurídica, fomentar buenas prácticas, facilitar la demostración de conformidad y, en último término, generar confianza en el despliegue responsable de la IA en un ámbito tan sensible como la justicia.

2.6.- Resolución del Parlamento Europeo de 6 de octubre de 2021 sobre la Inteligencia Artificial en el Derecho Penal y su Utilización por las Autoridades Policiales y Judiciales en Asuntos Penales (2020/2016(INI))

La Resolución del Parlamento Europeo de 6 de octubre de 2021 sobre la inteligencia artificial en el Derecho penal y su utilización por las autoridades policiales y judiciales en asuntos penales (2020/2016(INI)) es un instrumento jurídico no vinculante de singular relevancia en el proceso de configuración del marco regulatorio europeo en materia de inteligencia artificial (IA). Si bien carece de fuerza obligatoria *per se*, esta resolución erige un paradigma normativo que permeó necesariamente las presentes iniciativas legislativas de la Unión en este ámbito.

En primer término, es menester contextualizar la resolución en el marco del *acquis communautaire* y del debate legislativo en curso. Esta se inscribe en la estela de la entonces propuesta de reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas sobre la inteligencia artificial (Ley de Inteligencia Artificial), presentada por la

Comisión Europea el 21 de abril de 2021, hoy ya promulgado y en vigencia. En este sentido, la resolución objeto de análisis puede interpretarse como una toma de posición anticipada del Parlamento, que buscaba influir en la configuración final del texto legislativo.

El preámbulo de la resolución (considerandos A-M) establece el marco conceptual y axiológico que sustenta las subsiguientes disposiciones. Se reconoce, por una parte, el potencial transformador de la IA en términos de eficiencia y precisión (considerando A), al tiempo que se advierte sobre los *"enormes riesgos para los derechos fundamentales y las democracias basadas en el Estado de Derecho"* que esta tecnología puede entrañar. Esta dicotomía entre oportunidades y riesgos permea todo el texto de la resolución y refleja la complejidad del desafío regulatorio que la IA plantea.

Un aspecto crucial de la resolución es su anclaje en el acervo jurídico de la Unión en materia de derechos fundamentales. Se hace referencia explícita a la Carta de los Derechos Fundamentales de la Unión Europea, subrayando que los derechos allí consagrados deben garantizarse *"a lo largo de todo el ciclo de vida de la IA y las tecnologías conexas"* (considerando D). Esta invocación de la Carta no es meramente retórica, sino que tiene implicaciones jurídicas concretas, dado el carácter vinculante que le confiere el artículo 6 del Tratado de la Unión Europea.

La resolución aborda con particular énfasis la cuestión de la discriminación algorítmica (artículos 8-9). Se advierte sobre el riesgo de que los sistemas de IA perpetúen o exacerben sesgos preexistentes, con especial incidencia en grupos vulnerables o minorías étnicas. Esta preocupación se fundamenta en evidencia empírica que sugiere que muchas tecnologías de identificación basadas en algoritmos *"identifican y clasifican incorrectamente en un número desproporcionado de casos a las personas racializadas, a las personas pertenecientes a determinadas comunidades étnicas, a las personas LGTBI, a los niños y a las personas de edad avanzada, así como a las mujeres"* (artículo 9). La resolución vincula esta cuestión con el derecho a la no discriminación consagrado en el artículo 21 de la Carta, estableciendo así un nexo directo entre la regulación de la IA y la protección de los derechos fundamentales.

En materia de protección de datos personales, la resolución reafirma la plena aplicabilidad del Reglamento General de Protección de Datos (RGPD) y de la directiva sobre protección de datos en el ámbito penal (Directiva (UE) 2016/680) a los tratamientos de datos realizados mediante sistemas de IA. Esta reiteración es de suma importancia, pues subraya que la entonces futura legislación sobre IA no operará en un vacío jurídico, sino que debía integrarse armónicamente en

el marco normativo preexistente. Lo anterior evoca la jurisprudencia del Tribunal de Justicia de la Unión Europea en casos como Digital Rights Ireland (C-293/12)²³² y Tele2 Sverige (C-203/15)²³³, donde se establecieron límites estrictos a la retención masiva de datos personales por motivos de seguridad pública.

Un aspecto particularmente innovador y controvertido de la resolución radica en sus propuestas de prohibición o moratoria sobre determinadas aplicaciones de IA. Se propugna, inter alia, una prohibición de la "*puntuación de las personas a escala masiva mediante IA*" (artículo 32), práctica que se considera incompatible con los principios fundamentales de dignidad humana y no discriminación. Asimismo, se aboga por una moratoria en el despliegue de sistemas de reconocimiento facial para fines de identificación en espacios públicos, salvo para la identificación de víctimas de delitos "*hasta que las normas técnicas puedan considerarse plenamente acordes con los derechos fundamentales, los resultados obtenidos no estén sesgados y no sean discriminatorios, el marco jurídico prevea salvaguardias estrictas contra el uso indebido y un control y supervisión democráticos estrictos y existan pruebas empíricas de la necesidad y proporcionalidad del despliegue de estas tecnologías (...)*" (artículo 27). Estas propuestas de prohibición y moratoria representan un enfoque precautorio que privilegia la protección de los derechos fundamentales sobre consideraciones de eficacia policial o judicial.

La resolución dedica una atención considerable a la cuestión de la transparencia y la rendición de cuentas en el uso de sistemas de IA (artículos 17-18). Se aboga por que los algoritmos sean "*explicables, transparentes, trazables y comprobables*" (artículo 17), y se insta a las autoridades a hacer públicos los detalles de las herramientas de IA que utilizan, **incluyendo las tasas de falsos positivos y falsos negativos (artículo 33)**. Esta exigencia de transparencia se vincula directamente con el derecho a una buena administración consagrado en el artículo 41 de la Carta de los Derechos Fundamentales.

²³² Tribunal de Justicia de la Unión Europea. *Sentencia del Tribunal de Justicia (Gran Sala) de 8 de abril de 2014 en los asuntos acumulados C-293/12 y C-594/12, Digital Rights Ireland Ltd contra Minister for Communications, Marine and Natural Resources y otros y Kärntner Landesregierung y otros*. ECLI:EU:C:2014:238. Acceso el 25 de agosto de 2024. Recuperado de: <https://curia.europa.eu/juris/liste.jsf?num=293/12&language=es>

²³³ Tribunal de Justicia de la Unión Europea. *Sentencia del Tribunal de Justicia (Gran Sala) de 21 de diciembre de 2016 en los asuntos acumulados C-203/15 y C-698/15, Tele2 Sverige AB contra Post- och telestyrelsen y Secretary of State for the Home Department contra Tom Watson y otros*. ECLI:EU:C:2016:970. Acceso el 25 de agosto de 2024. Recuperado de: <https://curia.europa.eu/juris/liste.jsf?num=C-203/15&language=en>

En el ámbito procesal penal, la resolución subraya la necesidad de salvaguardar principios fundamentales como la presunción de inocencia y el derecho de defensa (artículo 10). Se hace especial hincapié en el derecho de las partes en un proceso penal a acceder a la información sobre el proceso de recopilación de datos y las evaluaciones realizadas mediante IA (artículo 14).

La resolución aborda también la cuestión de la responsabilidad jurídica por los daños causados por sistemas de IA (artículo 13). Se aboga por un "*régimen claro y justo para determinar la responsabilidad jurídica de las posibles consecuencias adversas derivadas de estas tecnologías digitales avanzadas*". Esta disposición anticipa uno de los aspectos más complejos y controvertidos de la regulación de la IA, a saber, cómo atribuir responsabilidad en un contexto de toma de decisiones algorítmicas.

De seguido, la resolución dedica una atención sustancial a la gobernanza de los sistemas de IA en el ámbito de la justicia penal, abogando por un enfoque basado en el riesgo (artículo 3). En este sentido, propugna la clasificación de los sistemas de IA utilizados en el ámbito policial y judicial como de "alto riesgo", lo cual conllevaría la aplicación de requisitos más estrictos en materia de transparencia, trazabilidad y supervisión humana. Esta categorización se alinea con el enfoque adoptado en la "*EU AI Act*", al establecer una presunción de alto riesgo para todas las aplicaciones de IA en este ámbito.

Un elemento central de este marco de gobernanza es la propuesta de realizar evaluaciones de impacto obligatorias sobre los derechos fundamentales antes de la implementación de cualquier sistema de IA en el ámbito policial o judicial (artículo 20). Esta disposición refleja una aplicación del principio de precaución en el ámbito de la IA, exigiendo una evaluación *ex ante* de los posibles riesgos para los derechos fundamentales. La resolución va más allá y aboga por auditorías periódicas obligatorias de todos los sistemas de IA utilizados por las autoridades policiales y judiciales (artículo 21), lo que implica un enfoque de supervisión continua que va más allá de la mera evaluación inicial.

La cuestión de la supervisión humana ocupa un lugar prominente en la resolución (artículo 16). Se enfatiza que:

“(...) en el contexto de las actividades judiciales y policiales, todas las decisiones con efectos legales deben ser tomadas siempre por un ser humano al que puedan pedirse cuentas de las decisiones adoptadas; considera que todas las personas objeto de sistemas de IA deben poder acceder a vías de recurso; recuerda que, en virtud del Derecho de la Unión, una persona tiene derecho a no ser objeto de una decisión que produzca efectos jurídicos que la conciernan o la afecte significativamente y que se base únicamente en el tratamiento automatizado de datos; subraya asimismo que la toma automatizada de decisiones individuales no debe basarse en las categorías especiales de datos personales, salvo que se hayan tomado medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado; destaca que el Derecho de la Unión prohíbe la elaboración de perfiles que dé lugar a la discriminación de personas físicas sobre la base de categorías especiales de datos personales; recuerda que las decisiones en el ámbito policial son casi siempre decisiones que tienen un efecto jurídico en la persona en cuestión, debido a la naturaleza ejecutiva de las autoridades policiales y sus acciones; destaca, en este sentido, que el uso de la IA puede influir en las decisiones humanas y afectar a todas las fases del procedimiento penal; considera, por tanto, que las autoridades que recurren a los sistemas de IA deben respetar unas normas jurídicas extremadamente estrictas y garantizar la intervención humana, especialmente cuando analicen los datos derivados de dichos sistemas; exige, por consiguiente, que se defiendan la discrecionalidad soberana de los jueces y las decisiones caso por caso; pide a la Comisión que prohíba el uso de la IA y las tecnologías conexas para proponer decisiones judiciales”, rechazando así la posibilidad de una toma de decisiones plenamente automatizada en el ámbito de la justicia penal”. (Énfasis agregado)

Este artículo merece un análisis pormenorizado dada su trascendencia para la configuración del futuro marco regulatorio de la IA en este sector.

En primer lugar, la disposición establece un principio fundamental: *“todas las decisiones con efectos legales deben ser tomadas siempre por un ser humano al que puedan pedirse cuentas de las decisiones adoptadas”*. Este principio tiene varias implicaciones jurídicas y prácticas de gran calado:

- **Responsabilidad jurídica:** al exigir que las decisiones sean tomadas por un ser humano "**al que puedan pedirse cuentas**", la resolución aborda directamente la cuestión de la responsabilidad jurídica en el contexto de la toma de decisiones asistida por IA. Esto plantea interrogantes sobre cómo se atribuirá la responsabilidad en casos donde la decisión humana se base en gran medida en recomendaciones de un sistema de IA.
- **Prohibición de la toma de decisiones plenamente automatizada:** esta disposición excluye efectivamente la posibilidad de sistemas de IA completamente autónomos en el ámbito de la justicia penal. Esto se alinea con el artículo 22 del RGPD, que establece el derecho a no ser objeto de decisiones basadas únicamente en el tratamiento automatizado que produzcan efectos jurídicos o afecten significativamente al interesado.
- **Naturaleza de la intervención humana:** la resolución no especifica el grado o la naturaleza de la intervención humana requerida. Esto deja abierta la cuestión de si una mera revisión formal de la recomendación de un sistema de IA sería suficiente, o si se requiere una evaluación sustantiva independiente por parte del ser humano.

La cuestión de la granularidad de la supervisión humana reviste particular importancia. Se plantea el interrogante de si dicha supervisión debe extenderse a cada etapa del proceso decisorio del sistema de IA o limitarse al escrutinio del resultado final. La primera opción podría resultar en una ralentización significativa del proceso judicial, mientras que la segunda podría ser insuficiente para detectar errores o sesgos algorítmicos subyacentes. Esta disyuntiva subraya la necesidad de establecer un equilibrio óptimo entre exhaustividad y eficiencia en la supervisión humana.

Asimismo, la implementación efectiva de la supervisión humana presupone la existencia de un cuerpo de supervisores dotados de las competencias técnicas necesarias para ejercer un control crítico sobre los sistemas de IA. Esta exigencia plantea desafíos significativos en términos de formación y selección del personal judicial, así como interrogantes sobre la posible necesidad de crear nuevos perfiles profesionales híbridos que combinen expertise jurídica y tecnológica.

En lo que concierne a la atribución de responsabilidad jurídica, la interposición de un supervisor humano en el proceso decisorio de un sistema de IA introduce complejidades adicionales. Se suscita la cuestión de cómo determinar la responsabilidad en casos donde una decisión, formalmente adoptada por un supervisor humano, pero sustancialmente basada en la

recomendación de un sistema de IA, resulte ser errónea o lesiva de derechos fundamentales. Este escenario pone de manifiesto las limitaciones del marco jurídico tradicional de responsabilidad en el contexto de sistemas de toma de decisiones híbridos humano-IA.

El artículo también hace referencia explícita al derecho de las personas "*a no ser objeto de una decisión que produzca efectos jurídicos que la conciernan o la afecte significativamente y que se base únicamente en el tratamiento automatizado de datos*". Esta disposición refuerza el principio establecido en el artículo 22 del RGPD y lo extiende específicamente al ámbito de la justicia penal. Es particularmente relevante dado que, como señala la resolución, "las decisiones en el ámbito policial son casi siempre decisiones que tienen un efecto jurídico en la persona en cuestión".

Un aspecto particularmente notable es el reconocimiento de que "*el uso de la IA puede influir en las decisiones humanas y afectar a todas las fases del procedimiento penal*". Esta afirmación reconoce el riesgo de lo que se conoce como "automatization bias", donde los seres humanos tienden a confiar excesivamente en las recomendaciones de los sistemas automatizados. Esto plantea la necesidad de desarrollar salvaguardias no solo contra la toma de decisiones plenamente automatizada, sino, también, contra la influencia indebida de los sistemas de IA en las decisiones humanas.

Sobre esto último, en primer término, es menester considerar la implementación de mecanismos de transparencia algorítmica que permitan a los operadores jurídicos comprender, en la medida de lo posible, los fundamentos lógicos subyacentes a las recomendaciones generadas por sistemas de IA. Esto podría materializarse a través de la exigencia legal de "explicabilidad" de los algoritmos utilizados en el ámbito judicial, requiriendo que estos sistemas sean capaces de proporcionar justificaciones claras y comprensibles de sus *outputs*. Paralelamente, se hace imperativo el desarrollo de programas de formación judicial especializados que doten a jueces y fiscales de las competencias necesarias para evaluar críticamente las recomendaciones algorítmicas, fomentando así una "alfabetización en IA" en el seno de la judicatura.

Secundariamente, se podría contemplar la introducción de mecanismos de "contra-análisis" o "segunda opinión" algorítmica, donde las recomendaciones de un sistema de IA sean

contrastadas sistemáticamente con las de otros sistemas o con análisis realizados por expertos humanos independientes. Este enfoque de "checks and balances" algorítmicos podría contribuir a mitigar el riesgo de una confianza excesiva en un único sistema de IA. Asimismo, la implementación de protocolos de "alerta de sesgo", que se activen cuando las decisiones humanas muestren una concordancia estadísticamente anómala con las recomendaciones de IA, podría servir como salvaguardia adicional contra la influencia algorítmica indebida. Estas medidas, en su conjunto, apuntan hacia la construcción de un ecosistema judicial donde la IA actúe como un auxiliar del razonamiento jurídico humano, sin llegar a suplantarla o distorsionarla indebidamente.

El artículo profundiza lo anterior al exigir que se defienda "*la discrecionalidad soberana de los jueces y las decisiones caso por caso*". Esta disposición busca preservar la autonomía judicial frente a la estandarización que podría resultar del uso generalizado de sistemas de IA. Sin embargo, plantea interrogantes sobre cómo conciliar esta discrecionalidad con el uso de herramientas de IA diseñadas para promover la consistencia en la toma de decisiones judiciales.

La conciliación entre la discrecionalidad judicial y la utilización de sistemas de inteligencia artificial (IA) en la toma de decisiones judiciales presenta un desafío jurídico-filosófico de singular complejidad. Por un lado, la discrecionalidad judicial, principio fundamental del Estado de Derecho, permite la adaptación de la norma abstracta a las particularidades del caso concreto, garantizando así una justicia individualizada. Por otro, los sistemas de IA prometen una mayor consistencia y previsibilidad en las decisiones judiciales, valores igualmente esenciales en un ordenamiento jurídico que aspira a la seguridad jurídica. Esta aparente antinomia entre flexibilidad y estandarización exige un análisis pormenorizado de sus implicaciones para la praxis judicial.

En este contexto, cabe preguntarse: ¿Cómo preservar el núcleo esencial de la función jurisdiccional -la interpretación y aplicación del Derecho al caso concreto- en un entorno tecnológico que tiende a la homogeneización decisoria? ¿Es posible articular un modelo de "discrecionalidad asistida" que aproveche las ventajas de la IA sin menoscabar la autonomía judicial? Estos interrogantes nos conducen a considerar la posibilidad de un paradigma híbrido, donde los sistemas de IA actúen como herramientas de apoyo, proporcionando al juez un marco de referencia basado en el análisis de casos precedentes y patrones jurisprudenciales, pero preservando siempre la facultad del juzgador para apartarse motivadamente de las sugerencias

algorítmicas. Tal enfoque requeriría, sin duda, un refinamiento de los protocolos de motivación judicial, exigiendo una explicación detallada no solo de la decisión final, sino, también, del proceso de valoración de las recomendaciones de IA y las razones para su eventual desestimación. De este modo, la discrecionalidad judicial no se vería coartada, sino enriquecida por un nuevo estrato de análisis, contribuyendo potencialmente a una jurisprudencia más reflexiva y matizada.

Finalmente, el artículo hace un llamamiento explícito a la Comisión para que "*prohíba el uso de la IA y las tecnologías conexas para proponer decisiones judiciales*". Esta es, quizás, la disposición más radical del artículo, ya que no solo prohíbe la toma de decisiones automatizada, sino incluso el uso de IA para proponer decisiones. Esto plantea interrogantes sobre la línea divisoria entre los sistemas de apoyo a la decisión, que podrían considerarse aceptables y los sistemas que "proponen" decisiones, que estarían prohibidos según esta disposición.

Y es que la demarcación entre sistemas de apoyo a la decisión y aquellos que "proponen" decisiones se revela como una frontera difusa, cuya delimitación precisa constituye un desafío hermenéutico de primer orden.

In limine, es menester dilucidar la *ratio legis* subyacente a esta prohibición. ¿Acaso pretende el legislador europeo salvaguardar la autonomía judicial frente a la injerencia algorítmica? ¿O se trata, más bien, de una manifestación del principio de precaución ante los riesgos, aún no plenamente comprendidos, de la delegación decisoria en sistemas artificiales? La respuesta a estos interrogantes no es baladí, pues de ella dependerá, en gran medida, la interpretación teleológica que se otorgue a la norma en cuestión.

Adentrándonos en la hermenéutica de la disposición, nos enfrentamos a la ardua tarea de deslindar conceptualmente los "sistemas de apoyo a la decisión" de aquellos que "proponen decisiones". Esta dicotomía, aparentemente diáfana, se torna nebulosa cuando la sometemos a un escrutinio riguroso. ¿Acaso no implica todo apoyo a la decisión, por su propia naturaleza, un elemento propositivo? ¿Dónde reside, pues, el límen que separa lo permisible de lo proscrito?

Para dilucidar esta cuestión, podríamos recurrir a la doctrina del acto administrativo en el Derecho administrativo. Así como se distingue entre los actos preparatorios y los actos definitivos de la administración, cabría establecer una diferenciación entre los sistemas de IA que realizan

funciones meramente preparatorias (recopilación y sistematización de información, análisis estadístico de precedentes) y aquellos que formulan propuestas decisorias concretas. Sin embargo, esta distinción, aparentemente nítida, se difumina ante la sofisticación creciente de los sistemas de IA.

En el Derecho administrativo, la distinción entre actos preparatorios y actos decisorios se fundamenta en la teoría de la formación de la voluntad administrativa. Los actos preparatorios, caracterizados por su naturaleza instrumental y no definitiva, se conciben como eslabones en la cadena procedural que conduce a la manifestación final de la voluntad administrativa. Por el contrario, los actos decisorios, dotados de eficacia jurídica inmediata, constituyen la cristalización de dicha voluntad y son susceptibles de impugnación directa.

Trasladando este esquema conceptual al ámbito de los sistemas de IA en la administración de justicia, podríamos establecer un paralelismo entre los sistemas que realizan funciones preparatorias y los actos administrativos de trámite. Estos sistemas, que se limitarían a la recopilación y sistematización de información, al análisis estadístico de precedentes o a la elaboración de resúmenes jurisprudenciales, no formularían *per se* una propuesta decisoria, sino que proporcionarían el sustrato informativo para la ulterior toma de decisión judicial.

En el otro extremo del espectro, los sistemas de IA que formulan propuestas decisorias concretas podrían equipararse, *mutatis mutandis*, a los actos administrativos resolutorios. Estos sistemas, al sugerir un fallo específico o una línea de argumentación jurídica determinada, estarían cruzando la línea que separa el mero apoyo de la proposición decisoria.

No obstante, esta analogía, aparentemente diáfana, se torna problemática ante la creciente sofisticación de los sistemas de IA. Consideremos, *verbi gratia*, un sistema que, si bien no formula explícitamente una propuesta de fallo, realiza un análisis predictivo de la probabilidad de éxito de diferentes argumentos jurídicos en un caso concreto. ¿Podría considerarse que tal sistema está realizando una función meramente preparatoria, o estaría *de facto* orientando la decisión judicial de manera tan significativa que equivaldría a una proposición implícita?

Consideremos, un sistema de IA que, si bien no formula explícitamente una propuesta de decisión, presenta la información de manera tan selectiva y estructurada que, *de facto*, está

orientando la decisión judicial en una dirección determinada. ¿Podría argumentarse que tal sistema está "proponiendo" una decisión de manera implícita? La respuesta a esta pregunta nos conduce a territorios jurídicos ignotos, donde los principios tradicionales de la interpretación normativa se revelan insuficientes.

Más aún, la prohibición en cuestión suscita interrogantes de calado constitucional. ¿Cómo conciliar esta restricción con el principio de independencia judicial consagrado en las constituciones de los Estados miembros y en la propia Carta de Derechos Fundamentales de la Unión Europea? ¿No podría argüirse que la prohibición del uso de determinadas herramientas tecnológicas constituye, en sí misma, una injerencia en la autonomía decisoria de los jueces?

Por otra parte, la implementación práctica de esta prohibición plantea desafíos técnicos y procedimentales de envergadura. ¿Cómo se supervisará el cumplimiento de esta norma? ¿Serán necesarios mecanismos de auditoría algorítmica para verificar que los sistemas de IA utilizados en el ámbito judicial no crucen la línea entre el apoyo y la proposición de decisiones? Estas cuestiones apuntan hacia la necesidad de desarrollar un nuevo corpus normativo y técnico en materia de gobernanza algorítmica judicial.

En última instancia, la analogía expuesta, si bien iluminadora, revela sus limitaciones ante la naturaleza *sui generis* de los sistemas de IA en el ámbito judicial. La fluidez y adaptabilidad de estos sistemas desafían las categorías jurídicas tradicionales, exigiendo la elaboración de nuevos marcos conceptuales que den cuenta de su complejidad. Quizás, más que buscar una analogía perfecta con doctrinas jurídicas preexistentes, el desafío radica en desarrollar un nuevo paradigma jurídico que reconozca la naturaleza única de los sistemas de IA en la administración de justicia. Este nuevo paradigma debería ser lo suficientemente flexible como para adaptarse a la rápida evolución tecnológica, pero, también, lo suficientemente robusto como para salvaguardar los principios fundamentales del Estado de Derecho.

Continuando con el análisis del resto del articulado, la resolución aborda también la cuestión de la formación y capacitación del personal policial y judicial en materia de IA (artículo 23). Se subraya la necesidad de una "*formación especializada considerable*" para garantizar una comprensión adecuada de los riesgos y limitaciones de los sistemas de IA. Esta disposición

reconoce implícitamente que la eficacia de cualquier marco regulatorio dependerá en última instancia de la capacidad de los operadores humanos para implementarlo correctamente.

Un aspecto particularmente controvertido de la resolución es su posición sobre el uso de tecnologías de reconocimiento facial y otras tecnologías biométricas (artículos 25-31). Se propone una prohibición del "*uso de análisis automatizados o el reconocimiento en espacios accesibles al público de otras características humanas, como los andares, las huellas dactilares, el ADN, la voz y otras señales biométricas y de comportamiento*" (artículo 26). Esta propuesta de prohibición amplia refleja una preocupación profunda por el potencial de estas tecnologías para facilitar una vigilancia masiva y vulnerar el derecho a la privacidad.

La resolución aborda también la cuestión del uso de bases de datos de reconocimiento facial privadas por parte de las autoridades policiales (artículo 28). Se expresa una "gran preocupación" por el uso de bases de datos como Clearview AI y se insta a la comisión a prohibir el uso de tales bases de datos en el ámbito de la garantía del cumplimiento de la ley. Esta disposición plantea cuestiones complejas sobre la interacción entre actores públicos y privados en el ámbito de la seguridad y la vigilancia.

Clearview AI es una empresa tecnológica estadounidense que ha desarrollado una controvertida base de datos de reconocimiento facial. Esta compañía merece una explicación detallada dado su papel central en el debate sobre la privacidad y la vigilancia en la era digital.

Esta empresa ha creado una base de datos masiva de imágenes faciales, que según informes, contiene más de 10 mil millones de fotografías²³⁴. Lo que hace particularmente polémica a esta base de datos es su método de recopilación: la empresa ha "raspado" (scraping) imágenes de redes sociales y otros sitios web públicos, a menudo sin el conocimiento o consentimiento explícito de las personas representadas en las imágenes o de las plataformas de las que se extraen.

El sistema de Clearview AI funciona de la siguiente manera: cuando se le proporciona una imagen facial, el software busca en su vasta base de datos y devuelve todas las imágenes coincidentes, junto con enlaces a las fuentes en línea donde se encontraron esas imágenes. Esto

²³⁴ Kashmir Hill, "The Secretive Company That Might End Privacy as We Know It," The New York Times, 18 de enero de 2020, <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>

permite, en teoría, identificar a una persona a partir de una sola fotografía y obtener información adicional sobre ella a través de las fuentes vinculadas

La empresa ha comercializado su tecnología principalmente a agencias de aplicación de la ley y otras entidades gubernamentales, argumentando que puede ser una herramienta valiosa para la investigación criminal y la seguridad nacional. Sin embargo, este uso ha suscitado graves preocupaciones entre los defensores de la privacidad y los derechos civiles.

Como se expuso en otros acápite, en el contexto de la Unión Europea, el uso de tecnología como Clearview AI es particularmente problemático dado el robusto marco de protección de datos establecido por el Reglamento General de Protección de Datos (RGPD). El RGPD establece estrictos requisitos para el procesamiento de datos biométricos, que incluyen una base legal clara y, en muchos casos, el consentimiento explícito del sujeto de los datos.

La mención específica de Clearview AI en la resolución del Parlamento Europeo refleja la creciente preocupación a nivel institucional en la UE sobre el uso de tales tecnologías. Al instar a la prohibición de estas bases de datos privadas en el ámbito de la aplicación de la ley, el Parlamento Europeo está tomando una posición firme en favor de la protección de la privacidad y contra la normalización de la vigilancia biométrica masiva.

La resolución dedica una atención considerable a la cuestión de la actuación policial predictiva (artículo 24). Se advierte sobre los riesgos de estos sistemas para perpetuar sesgos y discriminaciones y se cita la experiencia de varias ciudades de los Estados Unidos que han abandonado el uso de tales sistemas. Esta disposición refleja una preocupación más amplia sobre el uso de la IA para predecir comportamientos delictivos, una práctica que plantea interrogantes fundamentales sobre la presunción de inocencia y el libre albedrío.

La resolución aborda, también, la cuestión de la investigación y el desarrollo en el ámbito de la IA (artículo 35). Se insta a los Estados miembros a "apoyar y promover la investigación y desarrollo de AI solutions en apoyo de resultados sociales y ambientalmente beneficiosos". Esta disposición refleja un reconocimiento de que la regulación de la IA no debe obstaculizar la innovación, sino más bien canalizarla hacia fines socialmente beneficiosos.

Finalmente, la resolución hace un llamamiento a la comisión para que examine "si es necesaria una acción legislativa específica para especificar con más precisión los criterios y las condiciones para el desarrollo, el uso y el despliegue de aplicaciones y soluciones de IA por parte de las autoridades policiales y judiciales" (artículo 35). Lo anterior deja abierta la posibilidad de una legislación sectorial específica para el uso de la IA en el ámbito de la justicia penal, más allá del marco general establecido en la propuesta de reglamento de la comisión.

La resolución de marras representa un hito significativo en la evolución del marco regulatorio europeo sobre IA en el ámbito de la justicia penal. A través de sus disposiciones detalladas y a menudo controvertidas, establece un paradigma normativo que privilegia la protección de los derechos fundamentales y adopta un enfoque precautorio frente a los riesgos potenciales de la IA. Si bien su naturaleza no vinculante podría, *prima facie*, limitar su impacto directo, su valor como expresión de la voluntad política del Parlamento Europeo y como fuente de *soft law* no debe subestimarse. Las directrices y los principios enunciados en esta resolución proporcionan un marco conceptual y normativo que, previsiblemente, informó e informará el desarrollo legislativo y jurisprudencial en esta materia en los años venideros, configurando así el futuro *landscape* regulatorio de la IA en la Unión Europea.

2.7.- Libro Blanco sobre Inteligencia Artificial publicado por la Comisión Europea el 19 de febrero de 2020

2.7.1.- Naturaleza Jurídica y Ámbito de Aplicación

El Libro Blanco sobre la inteligencia artificial, promulgado por la Comisión Europea el 19 de febrero de 2020, es un instrumento de *soft law sui generis* en el ordenamiento jurídico de la Unión Europea. Su naturaleza jurídica, si bien carente de la fuerza vinculante propia del derecho positivo, trasciende la mera declaración programática para configurarse como un acto preparatorio de singular relevancia en el iter legislativo comunitario.

Desde una perspectiva dogmática, el Libro Blanco se inscribe en la categoría de los actos atípicos de la Comisión, cuya base jurídica, aunque no explícitamente consagrada en el artículo 288 del TFUE, encuentra su fundamento en el principio de atribución de competencias y en la

potestad de iniciativa legislativa que el artículo 17 del TUE confiere a la Comisión²³⁵. Esta atipicidad formal, lejos de menoscabar su relevancia jurídica, dota al documento de una flexibilidad y adaptabilidad particularmente idóneas para abordar los desafíos regulatorios que plantea una tecnología en constante evolución como la IA.

La eficacia normativa del Libro Blanco, si bien desprovista de la imperatividad característica de las fuentes tradicionales del derecho, se proyecta en múltiples dimensiones que merecen un análisis pormenorizado:

1. **Función pre-legislativa:** el documento actúa como un precursor normativo que prefigura el contenido material de la presente y futura legislación vinculante. En este sentido, el Libro Blanco opera como una suerte de "lex ferenda", anticipando y modelando el marco regulatorio de la IA en la UE. Esta función preparatoria se ha visto cristalizada en el Reglamento sobre IA analizado in extenso en la presente tesis, que recoge y desarrolla gran parte de los principios y conceptos esbozados en el Libro Blanco.
2. **Efecto hermenéutico:** los principios y conceptos delineados en el Libro Blanco están llamados a desempeñar un papel crucial en la interpretación teleológica y sistemática del futuro acervo comunitario en materia de IA. En este sentido, el documento puede ser invocado como un elemento de referencia en la labor hermenéutica del Tribunal de Justicia de la Unión Europea, en línea con la jurisprudencia establecida en el asunto Grimaldi³, que reconoce el deber de los órganos jurisdiccionales nacionales de tomar en consideración las recomendaciones de la comisión a la hora de interpretar las disposiciones nacionales o comunitarias²³⁶.
3. **Efecto de armonización indirecta:** a pesar de su carácter no vinculante, el Libro Blanco ejerce una influencia armonizadora *de facto* en las legislaciones nacionales, promoviendo

²³⁵ *1. La Comisión promoverá el interés general de la Unión y tomará las iniciativas adecuadas con este fin. Velará por que se apliquen los Tratados y las medidas adoptadas por las instituciones en virtud de éstos. Supervisará la aplicación del Derecho de la Unión bajo el control del Tribunal de Justicia de la Unión Europea. Ejecutará el presupuesto y gestionará los programas. Ejercerá asimismo funciones de coordinación, ejecución y gestión, de conformidad con las condiciones establecidas en los Tratados. Con excepción de la política exterior y de seguridad común y de los demás casos previstos por los Tratados, asumirá la representación exterior de la Unión. Adoptará las iniciativas de la programación anual y plurianual de la Unión con el fin de alcanzar acuerdos interinstitucionales.*

²³⁶ Tribunal de Justicia de la Unión Europea, Sentencia de 13 de diciembre de 1989, *Salvatore Grimaldi contra Fonds des maladies professionnelles*, C-322/88, ECLI:EU:C:1989:646, EUR-Lex, <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A61988CJ0322>

la convergencia normativa en un ámbito particularmente sensible a la fragmentación regulatoria. Este efecto se ve potenciado por el principio de cooperación leal consagrado en el artículo 4.3 del TUE²³⁷, que impone a los Estados miembros la obligación de abstenerse de adoptar medidas que puedan poner en peligro la realización de los objetivos de la Unión.

En cuanto a su ámbito de aplicación *ratione materiae*, el Libro Blanco adopta una concepción holística de la IA que abarca todo el ciclo de vida de estos sistemas. Esta aproximación omnicomprensiva se traduce en un enfoque regulatorio basado en el riesgo, que modula la intensidad de la intervención normativa en función del nivel de riesgo que presenta cada aplicación específica de la IA.

Ratione personae, el documento se dirige a un amplio espectro de destinatarios, incluyendo a los Estados miembros, las instituciones de la UE, la industria, la academia y la sociedad civil. Esta pluralidad de interlocutores refleja la voluntad de la comisión de adoptar un enfoque *multistakeholder* en la gobernanza de la IA, en consonancia con el principio de buena gobernanza enunciado en el Libro Blanco sobre la Gobernanza Europea²³⁸.

Cómo precisa el doctrinario español Moises Barrio:

“El Libro Blanco esboza el diseño de lo que califica como un ecosistema de confianza exclusivo en materia de IA, una meta que constituye un objetivo político en sí mismo, velando por el cumplimiento del acervo europeo en la materia, especialmente las normas de protección de los derechos fundamentales y los derechos de los consumidores, protegiendo en particular su seguridad frente a los desequilibrios informativos de la toma

²³⁷ Artículo 4. (...) 3. Conforme al principio de cooperación leal, la Unión y los Estados miembros se respetarán y asistirán mutuamente en el cumplimiento de las misiones derivadas de los Tratados. Los Estados miembros adoptarán todas las medidas generales o particulares apropiadas para asegurar el cumplimiento de las obligaciones derivadas de los Tratados o resultantes de los actos de las instituciones de la Unión. Los Estados miembros ayudarán a la Unión en el cumplimiento de su misión y se abstendrán de toda medida que pueda poner en peligro la consecución de los objetivos de la Unión.

²³⁸ Comisión Europea, "Libro Blanco sobre la Gobernanza Europea," EUR-Lex: Acceso al Derecho de la Unión Europea, última modificación el 21 de febrero de 2008. Recuperado de: <https://eur-lex.europa.eu/ES/legal-content/summary/white-paper-on-governance.html>

*de decisiones mediante algoritmos, en mayor medida con relación a los sistemas de inteligencia artificial que operan en la UE y presentan un riesgo elevado”.*²³⁹

2.7.2.- Aspectos Clave para la Implementación de IA en la Administración de Justicia

El Libro Blanco sobre IA esboza un marco conceptual y normativo que resulta de capital importancia para la implementación de sistemas automatizados de decisión basados en IA en la administración de justicia. Un análisis crítico de sus propuestas revela implicaciones jurídicas de hondo calado para nuestro objeto de estudio:

a) Enfoque basado en el riesgo:

El Libro Blanco propone una taxonomía del riesgo que clasifica las aplicaciones de IA en función de su potencial impacto en los derechos fundamentales y la seguridad. En este contexto, se sugiere que las aplicaciones de IA en la administración de justicia deberían considerarse, *a priori*, como de "alto riesgo". El documento establece que:

"Una aplicación de IA debe considerarse de riesgo elevado cuando presente la suma de los dos criterios siguientes:

- *En primer lugar, que la aplicación de IA se emplee en un sector en el que, por las características o actividades que se llevan a cabo normalmente, es previsible que existan riesgos significativos. [...]*
- *En segundo lugar, que la aplicación de IA en el sector en cuestión se use, además, de manera que puedan surgir riesgos significativos"*²⁴⁰

El sector de la administración de justicia se menciona, explícitamente, como uno de los ámbitos en los que es previsible que existan riesgos significativos, lo que implicaría la sujeción de las aplicaciones de IA en este contexto a requisitos legales más estrictos.

²³⁹ Moisés Barrio Andrés, dir., *El Reglamento Europeo de Inteligencia Artificial* (Valencia: Tirant lo Blanch, 2024), p. 27, ISBN 978-84-1071-304-8. Recuperado de: biblioteca.nubedelectura.com/cloudLibrary/ebook/show/9788410713048

²⁴⁰ Comisión Europea, *Libro Blanco sobre la inteligencia artificial: un enfoque europeo orientado a la excelencia y la confianza*, COM(2020) 65 final, Bruselas, 19 de febrero de 2020, p. 21. Recuperado de: <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52020DC0065>

b) Requisitos para Sistemas de IA de Alto Riesgo:

El Libro Blanco propone un conjunto de requisitos obligatorios para los sistemas de IA catalogados como de alto riesgo, que serían aplicables a las aplicaciones en el ámbito judicial:

1. **Datos de entrenamiento:** "*Es necesario adoptar las medidas necesarias para garantizar que, en lo que se refiere a los datos utilizados para entrenar los sistemas de IA, se respeten los valores y normas de la UE, concretamente con relación a la seguridad y la legislación vigente para la protección de los derechos fundamentales*"²⁴¹. Para el ámbito judicial, esto implica garantizar que los conjuntos de datos utilizados para entrenar los sistemas de IA sean representativos, libres de sesgos y respetuosos con la privacidad.
2. **Conservación de registros y datos:** "*Se necesitan requisitos con relación a la conservación de registros sobre la programación de algoritmos, los datos empleados para entrenar sistemas de IA de elevado riesgo y, en algunos casos, la conservación de los datos en sí mismos*". En el contexto judicial, esto se traduce en la implementación de mecanismos de trazabilidad que permitan reconstruir el proceso decisorio de los sistemas de IA.
3. **Información:** "*Debe informarse claramente a los ciudadanos de cuándo están interactuando con un sistema de IA y no con un ser humano*"²⁴². Para la administración de justicia, esto implica la necesidad de informar a las partes cuando se utilicen sistemas de IA en el proceso judicial.
4. **Solidez y exactitud:** "*Los sistemas de IA (y desde luego las aplicaciones de IA de riesgo elevado) deben ser técnicamente sólidos y exactos para ser fiables*"²⁴³. En el ámbito judicial, esto se traduce en la necesidad de garantizar la precisión y fiabilidad de los sistemas de IA utilizados en la toma de decisiones o en el apoyo a esta.
5. **Supervisión humana:** "*La supervisión humana ayuda a garantizar que un sistema de IA no socave la autonomía humana o provoque otros efectos adversos*"²⁴⁴. En el contexto de la administración de justicia, esto implica preservar el papel del juez como garante último

²⁴¹ Ibid, p. 23.

²⁴² Ibid, p. 24.

²⁴³ Ibid, p. 25.

²⁴⁴ Ibid.

de la decisión judicial, evitando una delegación completa de la función jurisdiccional en sistemas automatizados.

c) Gobernanza y Cumplimiento:

El Libro Blanco propone la creación de una estructura de gobernanza europea para la IA, que incluiría mecanismos de evaluación de conformidad *ex ante* para los sistemas de alto riesgo:

"Se requiere una estructura de gobernanza europea sobre IA en forma de un marco para la cooperación de las autoridades nacionales competentes, a fin de evitar la fragmentación de responsabilidades, incrementar las capacidades de los Estados miembros y garantizar que Europa se provea a sí misma de la capacidad necesaria para probar y certificar los productos y servicios provistos de IA "²⁴⁵.

En el contexto judicial, esto podría implicar la necesidad de certificar los sistemas de IA antes de su implementación en los tribunales, garantizando así su adecuación a los estándares de calidad, seguridad y respeto a los derechos fundamentales.

d) Responsabilidad:

El Libro Blanco aborda la necesidad de adaptar los marcos de responsabilidad civil a los desafíos planteados por la IA:

"Las características particulares de numerosas tecnologías de IA, como la opacidad («efecto caja negra»), la complejidad, la imprevisibilidad y un comportamiento parcialmente autónomo, pueden hacer difícil comprobar el cumplimiento de la legislación vigente de la UE sobre la protección de los derechos fundamentales e impedir su cumplimiento efectivo. Puede ser que las fuerzas y cuerpos de seguridad y las personas afectadas carezcan de los medios para comprobar cómo se ha tomado una decisión determinada con ayuda de la IA y, por consiguiente, si se han respetado las normas pertinentes. Las personas físicas y las personas jurídicas pueden enfrentarse a dificultades

²⁴⁵

*en el acceso efectivo a la justicia en situaciones en las que estas decisiones les afecten negativamente*²⁴⁶.

En el ámbito *sub examine*, esta cuestión adquiere especial relevancia en relación con la atribución de responsabilidad por decisiones erróneas o discriminatorias adoptadas con asistencia de sistemas de IA.

e) Consideraciones Éticas:

El Libro Blanco subraya la importancia de desarrollar e implementar la IA de manera ética y conforme a los valores fundamentales de la UE:

*"Teniendo en cuenta el enorme impacto que puede tener la inteligencia artificial en nuestra sociedad y la necesidad de que suscite confianza, resulta clave que la inteligencia artificial europea se asiente en nuestros valores y derechos fundamentales, como la dignidad humana y la protección de la privacidad"*⁹.

En el contexto de la administración de justicia, esto se traduce en la necesidad de salvaguardar principios como la independencia judicial, la igualdad de armas procesales o la presunción de inocencia en el diseño y despliegue de sistemas de IA.

Así, el Libro Blanco sobre IA establece un marco conceptual y normativo integral que sienta las bases para el desarrollo de un ecosistema regulatorio robusto para la implementación de la IA en la administración de justicia europea. Sus propuestas, que han sido en gran medida incorporadas al proyecto de reglamento sobre IA, configuran un paradigma regulatorio que busca equilibrar el fomento de la innovación tecnológica con la salvaguarda de los derechos fundamentales y los principios del Estado de Derecho.

2.8.- Análisis Comparativo con Marcos Regulatorios de otras Jurisdicciones Hegemónicas

Como ha quedado claro de la exposición realizada en este trabajo, en el convulso e incierto escenario tecnológico contemporáneo, la inteligencia artificial se erige como una de las fuerzas

²⁴⁶ Ibid, p.2.

motrices más poderosas del progreso y, a la vez, una de las fuentes más inquietantes de riesgo para los derechos fundamentales y la institucionalidad democrática. No extraña, pues, que la comunidad internacional se encuentre inmersa en un frenético afán regulatorio, encaminado a domesticar lo que algunos califican como la cuarta revolución industrial.

La Unión Europea, por su parte, ha dado pasos firmes al proponer el Reglamento Europeo sobre la IA (EU AI Act), marcando directrices precisas de transparencia, supervisión humana y responsabilidad. Sin embargo, para comprender en toda su magnitud las opciones regulatorias disponibles, resulta menester analizar las aproximaciones que, desde otras latitudes y con diferente bagaje histórico-político, se han puesto en práctica. Entre todas ellas, destacan los casos de Estados Unidos y China, dos potencias que, por su capacidad de innovación y su peso geopolítico, dictan buena parte del signo de la IA a escala planetaria.

La elección de Estados Unidos y China como referentes en este análisis no es casual ni arbitraria. Ambas naciones ostentan el liderazgo en materia de inversión en IA, concentración de talento investigador y número de patentes tecnológicas:

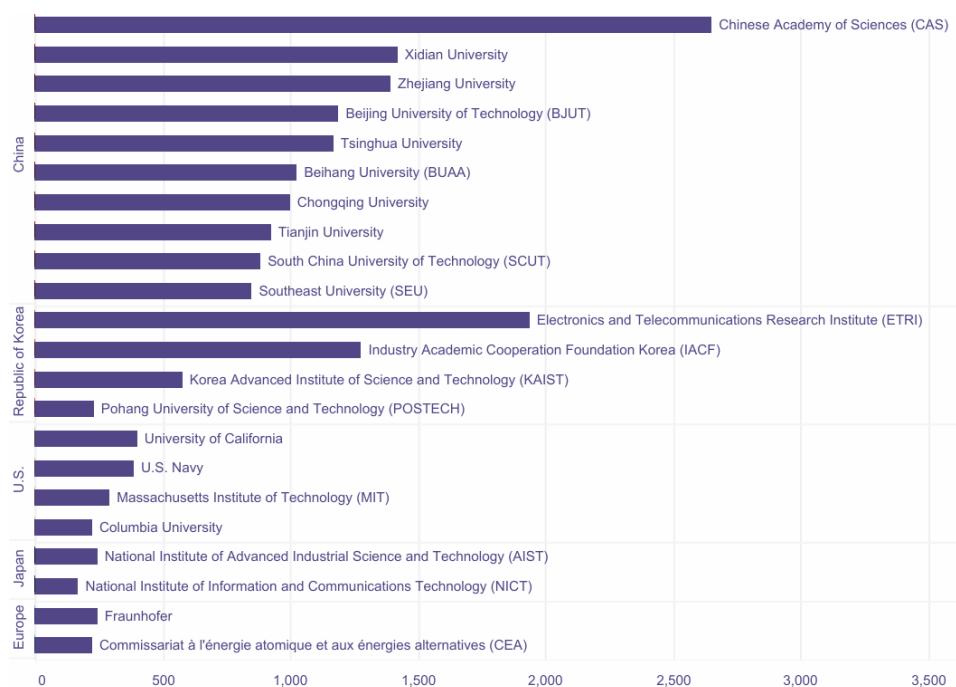


FIGURA 17: Principales solicitantes de patentes entre universidades y organizaciones públicas de investigación, clasificados por número de familias de patentes²⁴⁷.

²⁴⁷ WIPO, *WIPO Technology Trends 2019: Artificial Intelligence* (Ginebra: Organización Mundial de la Propiedad Intelectual, 2019). Recuperado de: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf

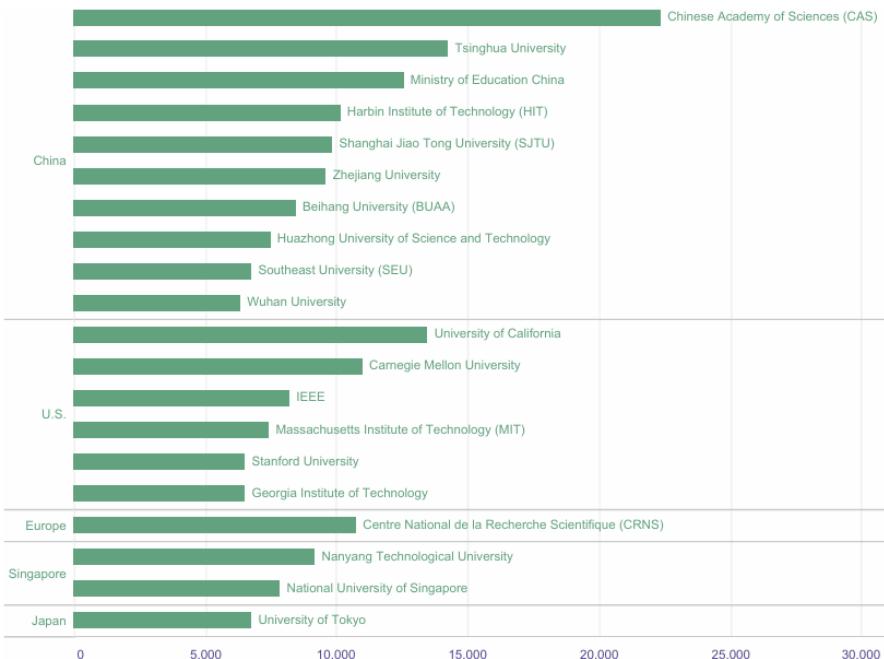


FIGURA 18: Las 20 principales universidades y organizaciones públicas de investigación que producen publicaciones científicas sobre inteligencia artificial, clasificadas por número de artículos. De las 20 principales organizaciones en publicaciones científicas sobre inteligencia artificial, 10 se encuentran en China, seis en Estados Unidos, dos en Singapur y una en Japón y Francia, respectivamente.

Asimismo, su influencia se expande más allá de sus fronteras: empresas estadounidenses y chinas dominan gran parte del mercado global de los servicios basados en IA; sus decisiones en torno a la protección de datos, la tutela de los derechos fundamentales o la transparencia algorítmica repercuten ineludiblemente en el resto del orbe.

2.8.1. Enfoque Regulatorio de Estados Unidos

La idiosincrasia estadounidense en el ámbito tecnológico hunde sus raíces en la amplia tradición de **libre mercado, desconfianza hacia la sobrerregulación** y preferencia por la **responsabilidad individual** antes que por la intervención estatal. El legado histórico del **laissez-faire** —heredado de la Revolución Industrial y reforzado por el auge de Silicon Valley— se traduce en la máxima de que la innovación florece sin trabas si no se ve constreñida por un **corsé normativo excesivo**.

Desde la década de 1980, con la expansión de la economía digital, esta visión liberal se ha consolidado bajo el axioma de que la tecnología es un **motor esencial** de la competitividad global. Por ende, la política pública busca minimizar la carga regulatoria, permitiendo que las **empresas emergentes** (start-ups), los **gigantes tecnológicos** (como Google o Microsoft) y diversos **laboratorios académicos** experimenten y asuman riesgos. La IA, considerada la piedra angular de la “revolución 4.0”, no escapa a este leitmotiv: se fomenta una autonomía amplia para desarrollar aplicaciones, dejando para las cortes y las agencias ejecutivas la tarea de moderar o corregir —*ex post*— posibles extralimitaciones.

Sin embargo, esta premisa liberal no significa ausencia total de normas. En un país de **federalismo complejo y gran diversidad estatal**, la adopción de reglas sectoriales, guías no vinculantes y ajustes *a posteriori* revela un equilibrio fino entre el interés nacional por liderar la IA y la necesidad de **evitar abusos** que socaven la confianza ciudadana.

A. Marco Institucional y Distribución de Competencias

- **La estructura Federal y la Iniciativa Presidencial**

En Estados Unidos, el **Gobierno Federal** interviene de manera puntual más que sistemática. Al carecer de un **Ministerio de Ciencia y Tecnología** centralizado, la coordinación en IA recae en diversas instancias:

- **La Casa Blanca** impulsa “grandes lineamientos” a través de órdenes ejecutivas y directrices de la Office of Science and Technology Policy (OSTP). Un ejemplo conspicuo fue la **Executive Order 13859**, “Maintaining American Leadership in AI”, que marcó un hito conceptual obstante, se trató de un instrumento de *soft law*, sin imposiciones firmes ni sanciones,
- **El Congreso** —salvo leyes sectoriales concretas— no ha promulgado, hasta ahora, una única “ley marco de IA”. En su lugar, ha aprobado la **National AI Initiative Act of 2020** para inyectar recursos a la investigación y crear estructuras consultivas, pero sin instaurar un régimen regulatorio rígido,
- **Las Agencias Ejecutivas** (ej.: la FDA, la FTC, el DOT o el CFPB) se encargan de **normar e inspeccionar** usos específicos de la IA en salud, transporte, finanzas, comercio, etc. Cada

agencia, con sus atribuciones legales y su tradición normativa, emite guías, avisos de reglamentación y cartas de cumplimiento.

El resultado práctico es un entramado relativamente **flexible y capaz de reconfigurarse** cuando la coyuntura lo exige. No hay un **órgano unificado** de supervisión ni un manual centralizado de IA; antes bien, cada esfera regula según sus prioridades, inspirada por el principio de que la competencia y la innovación no deben quedar asfixiadas.

- **Ámbito Estatal y Fragmentación de la Regulación**

A la complejidad federal se suma la **potestad de los Estados** para legislar en ámbitos como la protección del consumidor o la privacidad. Estados como California han promulgado leyes de vanguardia, como la **California Consumer Privacy Act (CCPA)** de 2018, que si bien no se diseñó específicamente para IA, impacta de lleno en la recolección y el tratamiento de datos que nutren estos sistemas²⁴⁸. Todo ello configura un panorama fragmentado, donde coexisten jurisdicciones estatales con diferentes grados de protección y exigencias.

La contrapartida de este enfoque es una notable fragmentación, que genera disparidades en la intensidad de la protección y en los requisitos de transparencia. Para un proveedor de IA, la localización geográfica y el sector de actividad definen qué pautas cumplir, en lugar de existir un único estándar nacional.

B. Principios Rectores y Documentos Clave

Aun con la ausencia de un “AI Act” federal, sí pueden identificarse textos clave que resumen la concepción liberal-estadounidense de la regulación de la IA.

- **Executive Order 13859 (2019): “Maintaining American Leadership in AI”²⁴⁹:**

Promulgada para “Mantener el Liderazgo Estadounidense en IA”, esta orden ejecutiva del

²⁴⁸ Ley de Privacidad del Consumidor de California de 2018 (California Consumer Privacy Act of 2018), Código Civil de California (Cal. Civ. Code), §§ 1798.100–1798.199.100, añadida por *Estatutos de 2018, Capítulo 55, Sección 3.* Recuperada de: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=CIV&division=3.&title=1.81.5.&part=4.&chapter=&article=

²⁴⁹ Orden Ejecutiva 13859 del 11 de febrero de 2019, *Mantener el Liderazgo Americano en Inteligencia Artificial*, publicada en el *Federal Register*, vol. 84, núm. 3967 (14 de febrero de 2019). Disponible en: <https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence>

presidente Trump buscó coordinar los recursos federales y propiciar un ambiente pro-innovación. Estableció líneas maestras: a) Aumentar la **inversión en I+D** de IA; b) Forjar directrices para el uso de la IA por agencias federales, minimizando la carga regulatoria innecesaria; c) Expandir la capacitación y difusión del conocimiento sobre IA. Lo notable es su énfasis en no “sofocar la innovación” y su apelación a la **autorregulación** empresarial. No entra en profundidades sobre derechos fundamentales ni obligaciones de transparencia algorítmica: **prima el incentivo a la competitividad**,

- **National AI Initiative Act of 2020²⁵⁰:** El **National AI Initiative Act** representa un paso legislativo más consistente, al crear oficinas y órganos consultivos específicos (National AI Initiative Office, National Artificial Intelligence Advisory Committee). Pretende coordinar esfuerzos federales e inversión en IA, pero tampoco fija un régimen completo de responsabilidad ni aborda con detalle la discriminación algorítmica,
- **The Blueprint for an AI Bill of Rights (OSTP, 2022)²⁵¹:** Bajo la administración Biden, la OSTP publicó el **Blueprint for an AI Bill of Rights**. Aunque su naturaleza es meramente declarativa, marca una **inflexión** conceptual hacia la protección del usuario. Sus cinco principios —sistemas seguros y efectivos, no discriminación, protección de la privacidad, aviso y explicación, y opciones humanas— reflejan una creciente conciencia sobre los riesgos de la IA. Por primera vez, un documento federal de alto nivel aboga por la equidad algorítmica (algorithmic fairness) y por “**human alternatives**” (supervisión o posibilidad de resolver sin IA). Sin embargo, su falta de obligatoriedad jurídica limita el impacto inmediato, quedando en las manos de las agencias su implementación efectiva.

²⁵⁰ **H.R.6216 - National Artificial Intelligence Initiative Act of 2020**, 116.^o Congreso (2019-2020), presentado el 12 de marzo de 2020 por la Representante Eddie Bernice Johnson (D-TX-30) y remitido al Comité de Ciencia, Espacio y Tecnología de la Cámara de Representantes. Recuperado de: <https://www.congress.gov/bill/116th-congress/house-bill/6216>

²⁵¹ **Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People**, Oficina de Política de Ciencia y Tecnología de la Casa Blanca (OSTP), octubre de 2022. Disponible en: <https://www.whitehouse.gov/ostp/ai-bill-of-rights>

C. Materialización del Liberalismo en la Práctica

El **enfoque liberal** estadounidense se cristaliza en varios puntos concretos, que ilustran su pragmatismo:

1. **No existe un RGPD local:** la protección de datos se rige por un mosaico de leyes sectoriales (HIPAA en salud, Gramm-Leach-Bliley en finanzas, COPPA para niños, etc.). Consecuencia: la IA dispone de mayor **libertad de recolección de datos** mientras no infrinja las contadas prohibiciones legales,
2. **Responsabilidad *a posteriori*:** cuando un sistema de IA causa perjuicios (discriminación en el crédito, sesgo racial en evaluaciones penales), el afectado recurre a la vía judicial o a la agencia competente. No hay una revisión previa obligatoria, salvo en sectores muy sensibles (ej.: FDA para IA médica),
3. **Intervención limitada en la toma de decisiones judiciales:** aunque existen algoritmos de valoración de riesgo (COMPAS en criminal), su uso no es uniforme ni está vetado federalmente. Varios estados lo adoptan para recomendar fianzas o penas, sin un mandato nacional que asegure transparencia. Los tribunales pueden aceptar o rechazar su uso conforme a la doctrina probatoria.
4. **Fomento de la colaboración público-privada:** el gobierno financia la investigación en IA (National Science Foundation, DARPA) e incentiva la participación de big tech en consorcios, confiando en la “buena fe” y en la “autorregulación responsable”²⁵²

En suma, la lógica liberal invita a la **autorregulación**, guiada por el temor a frenar la competitividad global del país. No se ignoran los peligros del sesgo algorítmico, pero se prioriza la intervención correctiva puntual sobre la imposición de exigencias *ex ante* de amplia envergadura.

²⁵² Melissa Heikkilä, "How's AI Self-Regulation Going? One Year on from the White House's Voluntary Commitments on AI," MIT Technology Review, 23 de julio de 2024. Recuperado de: https://www.technologyreview.com/2024/07/23/1095218/hows-ai-self-regulation-going/?utm_source=chatgpt.com

D. Críticas y Retos dentro de la Visión Liberal

Este paradigma anglosajón, si bien aplaudido por el sector privado, afronta una serie de objeciones insoslayables:

1. **Falta de uniformidad:** la coexistencia de normas estatales y sectoriales complica la seguridad jurídica de los desarrolladores de IA.
2. **Déficit de protección de datos:** a ojos comparados (particularmente de la UE), la ausencia de una ley federal de privacidad dificulta la salvaguarda de los derechos digitales.
3. **Vacíos de responsabilidad:** ¿quién responde si un sistema de IA produce discriminación? ¿La agencia que lo emplea, el programador, la empresa? La jurisprudencia no está unificada, y las demandas sobre sesgo no siempre prosperan por falta de tipificación clara.
4. **Transparencia limitada:** pese a la retórica del Blueprint for an AI Bill of Rights, no se ha articulado un mandato que obligue a empresas a revelar la lógica algorítmica de sistemas cruciales, ni existen procedimientos de auditoría universal.

El caso **COMPAS** (herramienta utilizada en sentencias penales en varios estados) exemplifica el choque: ProPublica expuso sesgos raciales en las puntuaciones de reincidencia²⁵³. Sin embargo, al ser un sistema amparado por derechos de autor, la compañía dueña del algoritmo se negó a revelar su metodología. La doctrina no ha establecido la obligación de divulgación total, representando un vacío que atenta contra la transparencia del procedimiento penal.

E.- Hacia un Posible Viraje o Consolidación

Pese a su impronta liberal, el “clima regulatorio” en Estados Unidos podría experimentar cambios, atendiendo a varios factores:

²⁵³ Julia Angwin, Jeff Larson, Surya Mattu, y Lauren Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks," ProPublica, 23 de mayo de 2016. Recuperado de: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

- **Presiones sociales:** ante la proliferación de casos de discriminación algorítmica o la manipulación masiva de datos, sectores académicos y organizaciones civiles claman por un estatuto unificado que salvaguarde derechos básicos²⁵⁴,
- **Competencia internacional:** el liderazgo de la Unión Europea con su AI Act (de innegable vocación extraterritorial) podría “arrastrar” a EE. UU. a adoptar normas más rigurosas, especialmente si las empresas estadounidenses desean seguir operando sin trabas en el mercado europeo²⁵⁵,
- **Iniciativas legislativas emergentes:** varios proyectos del Congreso, como el “Algorithmic Accountability Act”²⁵⁶, promueven la obligación de auditorías, la rendición de cuentas y la supervisión federal de herramientas IA. Aun con un avance lento, su sola existencia indica cierta voluntad de regular.

La disyuntiva radica entre un reforzamiento de la tradición liberal o un progresivo alineamiento con la Unión Europea en la imposición de obligaciones *ex ante* más detalladas. Lo más plausible es un **modelo híbrido**, con medidas sectoriales reforzadas y un mayor hincapié en el control de la discriminación, sin llegar a la uniformidad estricta que caracteriza al RGPD europeo.

F.- Conclusión: Un liberalismo que Busca Conciliar Innovación y Salvaguarda de Derechos

En definitiva, el **enfoque regulatorio de Estados Unidos** respecto de la IA surge como **liberal**, no solo por su vocación de fomentar la innovación y la competitividad global, sino, también, por su reticencia a promulgar un marco unificado de carácter vinculante. El Estado se erige como facilitador, fiándose de la interacción entre agencias, las directrices presidenciales y la disciplina de mercado para ir **corrigiendo disfunciones**. Esta filosofía parte de la convicción de

²⁵⁴ Jesse Bedayn, "Attempts to Regulate AI's Hidden Hand in Americans' Lives Flounder in US Statehouses," Associated Press, 23 de mayo de 2024. Recuperado de: https://apnews.com/article/artificial-intelligence-bias-discrimination-regulation-ai-ff1d0860663723079aac3666b38f2320?utm_source=chatgpt.com

²⁵⁵ KPMG. *How the EU AI Act Affects US-Based Companies: A Guide for CISOs and Other Business Leaders*. KPMG LLP, marzo de 2024. Recuperado de: <https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2024/decoding-eu-ai-act.pdf>

²⁵⁶ H.R. 5628, Algorithmic Accountability Act of 2023, 118th Congress (2023-2024), presentado por la representante Yvette D. Clarke [D-NY-9] el 21 de septiembre de 2023, remitido al Subcomité de Innovación, Datos y Comercio el 22 de septiembre de 2023,

que la excesiva normación podría ahogar la dinámica creativa de las start-ups y las grandes tecnológicas.

Sin embargo, las **presiones sociales** y la evidencia de que la IA puede conculcar derechos fundamentales impulsan, gradualmente, una demanda de mayor transparencia y responsabilidades definidas. El sistema estadounidense es versátil y se adapta a la coyuntura, por lo que no resulta descabellado imaginar un escenario a medio plazo en el cual el principio liberal coexista con normas de transparencia, rendición de cuentas y prohibición de prácticas discriminatorias, en la línea del **Blueprint for an AI Bill of Rights**.

Mientras tanto, en contraste con la aproximación europea, más holística y de alto nivel, el **modelo estadounidense** preserva su esencia liberal y fragmentada, confirmando el carácter histórico de un país donde la **libertad de emprendimiento** se prioriza por encima de la **intervención regulatoria**, a la espera de que las tensiones y controversias terminen por delinear un consenso legislativo más robusto.

2.8.2.- Modelo Regulatorio de la República Popular de China

La República Popular China ha emprendido, en los últimos lustros, un ascenso vertiginoso hasta erigirse como uno de los polos globales más poderosos en materia de inteligencia artificial (IA). El país ha logrado concentrar enormes capitales de inversión, aglutinar talento científico y alinear la acción de las grandes corporaciones tecnológicas con las directrices trazadas por el Partido Comunista Chino. Este rápido florecimiento de la IA se enmarca en un modelo normativo autoritario, que combina la planificación centralizada, la estrecha supervisión estatal y una retórica oficial que enfatiza la estabilidad social como justificación para un uso extensivo de la tecnología.

La aproximación de China a la IA, en consecuencia, difiere de forma radical de la tradición liberal de Estados Unidos, al tiempo que contrasta con el enfoque garantista y basado en derechos fundamentales que postula la Unión Europea. En las siguientes páginas se profundiza en la arquitectura legal y en los textos estratégicos más sobresalientes, exponiendo la lógica que subyace a la centralidad del Estado en la regulación de la IA y explorando sus implicaciones para la gobernanza tecnológica y el orden jurídico internacional.

A. Antecedentes y Planificación Estratégica

La República Popular China cimentó su estrategia en IA a partir de un documento seminal: el **New Generation Artificial Intelligence Development Plan²⁵⁷**, publicado por el Consejo de Estado en julio de 2017. Este plan, que aspira a posicionar a China como líder mundial en IA para 2030, articula metas escalonadas:

- **Para 2020:** la primera etapa, cuyo objetivo se fijó para el 2020, se centró en sincronizar el desarrollo tecnológico e industrial de la IA en China con los estándares avanzados a nivel global. En esta fase, se alcanzaron avances significativos en áreas como la inteligencia basada en grandes datos (big data), la inteligencia de medios cruzados (cross-media), la inteligencia de enjambre, la inteligencia híbrida aumentada y los sistemas autónomos inteligentes. Estos logros debían fortalecer las bases teóricas y tecnológicas necesarias para la innovación continua. Además, se establecieron estándares tecnológicos preliminares, sistemas de servicio y cadenas industriales que posicionaron a la industria de la IA como un motor clave del crecimiento económico, con ingresos proyectados en 150 mil millones de yuanes en sectores centrales y 1 billón de yuanes en áreas relacionadas. Paralelamente, se promovió un entorno propicio para la innovación mediante aplicaciones prácticas en áreas estratégicas y la atracción de talento de alto nivel, acompañado de un marco inicial de regulación y ética para guiar el desarrollo de la IA,
- **Para 2025:** en la segunda etapa, planificada para culminar en 2025, el énfasis recae en posicionar a China como líder mundial en el ámbito de la IA, tanto en investigación teórica como en aplicaciones prácticas. Se busca consolidar un sistema técnico avanzado que incorpore capacidades de aprendizaje autónomo, así como lograr avances significativos en sectores como la manufactura, la atención médica, las ciudades inteligentes y la defensa nacional. Esta fase proyecta un crecimiento industrial sustancial, con una industria central de IA valorada en más de 400 mil millones de yuanes e industrias relacionadas que alcanzarán los 5 billones de yuanes. Además, se contempla la creación de un sistema inicial de regulaciones, políticas y principios éticos, así como mecanismos para evaluar y

²⁵⁷ Consejo de Estado de la República Popular China, *Aviso del Consejo de Estado sobre la Impresión y Distribución del Plan de Desarrollo de la Nueva Generación de Inteligencia Artificial* (Guo Fa [2017] No. 35), emitido el 8 de julio de 2017 y publicado oficialmente el 20 de julio de 2017. Recuperado de: https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

gestionar la seguridad de las aplicaciones de IA, consolidando un marco institucional robusto que permita su expansión segura y controlada,

- **Para 2030:** la tercera etapa, prevista para 2030, establece la meta de posicionar a China como el principal centro de innovación en IA a nivel mundial. En esta fase, se espera que el país alcance un liderazgo absoluto en la investigación teórica, las aplicaciones tecnológicas y el desarrollo industrial. Los avances esperados incluyen progresos significativos en áreas como la inteligencia similar al cerebro, la inteligencia autónoma, la inteligencia híbrida y la inteligencia de enjambre, consolidando a China como un referente global en el ámbito científico y tecnológico. En términos económicos, la escala de la industria central de IA superará 1 billón de yuanes, mientras que las industrias relacionadas alcanzarán los 10 billones de yuanes, impulsando un ecosistema integral que abarque tecnologías clave, plataformas de soporte y aplicaciones inteligentes. Asimismo, se proyecta la formación de bases de innovación tecnológica de clase mundial y la implementación de un marco regulatorio y ético completo que garantice un desarrollo seguro y alineado con principios responsables.

A diferencia de otros países donde las directrices estratégicas provienen, principalmente, de actores privados, aquí el **órgano gubernamental** diseña y supervisa el cumplimiento de objetivos, conjuga a las grandes empresas y moviliza recursos financieros considerables. Esto responde a una lógica **estatista**: la IA sirve a la "grandeza nacional" y a la "prosperidad socialista", quedando siempre supeditada a la misión histórica del Partido Comunista.

B. Principales Instrumentos Jurídicos: Data Security Law y PIPL

Tras delinejar la visión macro, el paso siguiente fue dotar a este plan de cimientos legales. Así, la República Popular China promulgó dos leyes fundamentales: la **Ley de Seguridad de Datos (Data Security Law, 2021)** y la **Ley de Protección de la Información Personal (Personal Information Protection Law, 2021)**.

1. **Ley de Seguridad de Datos (Data Security Law)²⁵⁸**: Aprobada en junio de 2021, estableció un régimen de clasificación de los datos según su relevancia para la seguridad nacional y el desarrollo socioeconómico. El texto impone a las entidades que manejen datos "importantes" (o "core data") la adopción de salvaguardas reforzadas y el cumplimiento de procedimientos de autorización para su exportación transfronteriza. Aunque no se remite explícitamente a la IA, su efecto sobre la misma es profundo, pues la construcción de sistemas de inteligencia artificial a gran escala depende, en gran medida, de la disponibilidad de conjuntos masivos de datos.
2. **Ley de Protección de la Información Personal (PIPL)²⁵⁹**: En vigor desde noviembre de 2021, la PIPL constituye el primer instrumento general en China para la tutela de la privacidad de las personas. Inspirada en ciertos aspectos del RGPD europeo, introduce el principio de consentimiento, la minimización de datos y la prohibición de tratar datos sensibles sin justificación. Sin embargo, a diferencia del RGPD, otorga al Estado facultades para exigir la provisión de datos en casos de seguridad nacional o interés público. Es jurídicamente relevante señalar que estas facultades no constituyen poderes discrecionales amplios, sino que están: a) Vinculadas al cumplimiento de funciones estatutarias específicas b) Sujetas a principios de necesidad y proporcionalidad c) Enmarcadas en procedimientos y salvaguardas definidos d) Sometidas a obligaciones de protección de datos. De esta forma, la PIPL equilibra la protección de la privacidad individual con la potestad soberana del Partido para recabar y explotar datos en aras de la estabilidad social.

La conjunción de ambas leyes dota al **Partido-Estado** de un poder singular: supervisar el flujo de datos en cada etapa del ciclo de vida de la IA, **restringir** o **autorizar** su transferencia a empresas o a entidades extranjeras y **monitorear** la naturaleza y finalidad de los tratamientos algorítmicos.

²⁵⁸ Comité Permanente de la Asamblea Popular Nacional de la República Popular de China. Ley de Seguridad de los Datos de la República Popular China. 10 de junio de 2021. Recuperada de: <https://www.chinalawtranslate.com/en/datasecuritylaw/>

²⁵⁹ Comité Permanente de la Asamblea Popular Nacional, *Ley de Protección de la Información Personal de la República Popular China*, adoptada en la 30.^a reunión del Comité Permanente de la XIII Asamblea Popular Nacional el 20 de agosto de 2021 y promulgada mediante la Orden Presidencial N.^o 91, vigente a partir del 1 de noviembre de 2021. Recuperada de: <https://www.china-briefing.com/news/the-prc-personal-information-protection-law-final-a-full-translation/>

C. Normas Especiales sobre IA y la “Gestión de la Síntesis Profunda”

Además de las leyes generales, China ha promulgado reglas específicas para afrontar los desafíos de las tecnologías emergentes de IA, como el **deep learning**, la **síntesis profunda** y el **reconocimiento facial**. Entre las regulaciones más destacadas:

- **Provisiones sobre la Gestión de la Síntesis Profunda (CAC, 2023)**²⁶⁰: la Administración del Ciberespacio de China (CAC) emitió estas disposiciones para regular los llamados servicios de “deepfake” y la generación de contenido multimedia sintético. Exigen a los proveedores que se aseguren de que su tecnología no sea usada con fines ilícitos (manipulación política, difamación, pornografía) y que etiqueten apropiadamente los contenidos generados. El acento recae en **proteger la estabilidad social** y prevenir la diseminación de material que pudiera atentar contra la narrativa oficial,
- **Disposiciones sobre la Gestión de Recomendaciones Algorítmicas en Servicios de Información de Internet**²⁶¹: fundamentadas en el artículo 1 sobre la base del ordenamiento jurídico conformado por las leyes de Ciberseguridad, Seguridad de Datos y Protección de Información Personal, establecen en su artículo 2 su ámbito de aplicación territorial sobre el territorio continental chino, regulando específicamente el uso de tecnologías algorítmicas para la provisión de servicios de información. El marco normativo instituye, mediante los artículos 6 al 15, un régimen de obligaciones sustantivas para los proveedores, incluyendo la implementación de sistemas de gestión de seguridad algorítmica (artículo 7), la prohibición de modelos algorítmicos que induzcan adicción (artículo 8), y el establecimiento de mecanismos de intervención manual y selección independiente de usuarios (artículo 11), complementado por un sistema de protección de derechos consagrado en los artículos 16 al 22, que contempla específicamente salvaguardas para menores (artículo 18) y adultos mayores (artículo 19). Implementa, a través de los

²⁶⁰ Administración del Ciberespacio de China, Ministerio de Industria y Tecnología de la Información y Ministerio de Seguridad Pública, Provisions on the Administration of Deep Synthesis Internet Information Services. Promulgado el 25 de noviembre de 2022. Recuperado de: <https://www.chinalawtranslate.com/en/deep-synthesis/>

²⁶¹ Oficina de Administración del Ciberespacio de China, Ministerio de Industria y Tecnología de la Información, Ministerio de Seguridad Pública y Administración Estatal para la Regulación del Mercado, Disposiciones sobre la Gestión de Recomendaciones Algorítmicas en los Servicios de Información de Internet, promulgada el 31 de diciembre de 2021, vigente a partir del 1 de marzo de 2022. Recuperado de: <https://www.chinalawtranslate.com/en/algorithms/>

artículos 23 al 29, un sistema de supervisión estatal jerárquico y categorizado, estableciendo en el artículo 24 la obligación de registro para proveedores con capacidad de movilización social dentro de los 10 días hábiles posteriores al inicio de sus servicios. El régimen sancionatorio, codificado en los artículos 31 al 33, contempla sanciones graduales que van desde advertencias hasta multas de 10.000 a 100.000 RMB, incluyendo la posibilidad de suspensión temporal de actualizaciones de información. Este marco regulatorio se complementa con disposiciones específicas sobre evaluaciones de seguridad (artículo 27) y obligaciones de cooperación con las autoridades supervisoras (artículo 28, párrafo 2), configurando así un sistema integral de gobernanza algorítmica que entró en vigor el 1 de marzo de 2022, según establece el artículo 35,

- **Medidas Provisionales para la Gestión de los Servicios de Inteligencia Artificial Generativa²⁶²:** instrumenta un régimen regulatorio tripartito que comprende: (i) un sistema de control previo, manifestado en la obligación de los proveedores de realizar evaluaciones de seguridad y registros algorítmicos para servicios con capacidad de movilización social (Art. 17); (ii) un control concurrente, materializado en la obligación de implementar medidas efectivas para incrementar la transparencia y confiabilidad del contenido generado (Art. 4.5), así como en el deber de establecer mecanismos de quejas y denuncias con portales accesibles (Art. 15); y (iii) un control posterior, ejercido mediante la facultad de los departamentos competentes para realizar inspecciones de supervisión sobre los servicios de IA generativa (Art. 19). El marco normativo establece un sistema de obligaciones específicas para los proveedores que incluye: la obtención del consentimiento para el tratamiento de información personal (Art. 7.3), el etiquetado del contenido generado (Art. 12), la implementación de medidas contra contenidos ilegales (Art. 14), y la protección de datos de usuarios (Art. 11). El régimen sancionador se estructura en tres niveles: (i) administrativo general, que contempla advertencias, críticas públicas y órdenes de corrección; (ii) administrativo especial, que puede resultar en la suspensión de servicios; y (iii) penal, aplicable cuando las infracciones constituyan delitos (Art. 21).

²⁶² Oficina de Administración del Ciberespacio de China, Comisión Nacional de Desarrollo y Reforma, Ministerio de Educación, entre otros, *Medidas Provisionales para la Gestión de los Servicios de Inteligencia Artificial Generativa*, promulgadas el 10 de julio de 2023, vigentes desde el 15 de agosto de 2023. Disponible en http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

Adicionalmente, se implementa un sistema de control transfronterizo que faculta al departamento estatal de información en Internet para adoptar medidas técnicas contra servicios provistos desde el exterior que incumplan la normativa (Art. 20), complementado con requisitos específicos para la inversión extranjera en servicios de IA generativa (Art. 23),

- **Proyecto de Ley de Inteligencia Artificial de la República Popular de China (2024)²⁶³:**
En febrero de 2024, se circuló un borrador preliminar de la "Ley de Inteligencia Artificial de la República Popular China" entre académicos y expertos legales. Establece un marco regulatorio multinivel para el desarrollo, provisión y utilización de sistemas de inteligencia artificial. La propuesta legislativa articula principios fundamentales como la centralidad humana, el desarrollo innovador y la supervisión prudencial, instituyendo obligaciones diferenciadas para desarrolladores, proveedores y usuarios, con especial énfasis en la regulación de la IA crítica, definida como aquella que impacta significativamente en derechos fundamentales o infraestructura crítica. Desde una perspectiva sistemática, el texto normativo establece un régimen de supervisión estatal coordinado que comprende mecanismos de evaluación de riesgos, requisitos de registro y certificación, así como un sistema gradual de responsabilidad civil y administrativa. Se destaca la incorporación de disposiciones específicas sobre modelos fundacionales y escenarios de aplicación especial, como el uso judicial o médico, junto con un marco de cooperación internacional que enfatiza la participación china en la gobernanza global de la IA, todo ello complementado con un régimen sancionador que contempla multas de hasta 50 millones de yuanes o el 5 % de la facturación anual para infracciones graves. En materia de inteligencia artificial aplicada al ámbito judicial, el artículo 70 del proyecto normativo establece un régimen jurídico específico que se fundamenta en cinco principios cardinales: la seguridad y legalidad como presupuestos operativos básicos; la equidad e

²⁶³ East China University of Political Science and Law, Digital Rule of Law Institute, "Artificial Intelligence Law of the People's Republic of China (Draft for Suggestions from Scholars) [中华人民共和国人工智能法 (学者建议稿)]." Traducido por Etcetera Language Group, Inc., editado por Ben Murphy. CSET Translation Manager, 2 de mayo de 2024.. Recuperado de: https://cset.georgetown.edu/publication/china-ai-law-draft/?utm_source=chatgpt.com

imparcialidad como garantías procesales fundamentales; la función auxiliar en el proceso decisario judicial; la transparencia y credibilidad como elementos de legitimación; y la conformidad con el orden público y la moral como límites axiológicos de su implementación. La disposición normativa instituye dos salvaguardas procesales fundamentales que delimitan el alcance de la inteligencia artificial en sede judicial: en primer término, circscribe el valor jurídico de los pronunciamientos generados mediante sistemas de IA a un carácter meramente referencial, preservando así la autonomía decisoria del órgano jurisdiccional; y en segundo lugar, consagra una garantía procesal en favor de los justiciables, materializada en el derecho a retirarse de la interacción con productos y servicios de IA en cualquier momento del proceso, salvaguardando así el derecho fundamental al debido proceso y la tutela judicial efectiva.

D. Lógica Centralizada de la Gobernanza: Confucionismo Digital

La lógica centralizada que subyace a la gobernanza de la inteligencia artificial en la República Popular China hunde sus raíces en los principios cardinales del confucianismo tradicional, cuya influencia permea el ethos político y social del país incluso en la era digital. Esta cosmovisión, que algunos académicos han acuñado como "confucianismo digital"²⁶⁴, se erige sobre pilares axiológicos como la armonía social, la piedad filial y la primacía del colectivo sobre el individuo, nociones que han sido reinterpretadas y adaptadas por el Partido Comunista Chino para cimentar un modelo de gobernanza tecnológica sin parangón en el orden jurídico internacional.

En el núcleo de este paradigma regulatorio se sitúa el rol preponderante del Estado y del Partido como guardianes del interés común y garantes últimos del orden social. Así como en la tradición confuciana el poder político asume una función rectora en la preservación de la estabilidad y la cohesión del cuerpo social, en el ámbito de la IA el Partido-Estado se arroga la potestad de supervisar, encauzar y, cuando lo estima necesario, restringir el desarrollo y despliegue

²⁶⁴ Culhong Cai y Jiahui Yin, "Cultural and Ethical Foundations of AI Governance Divergence: A Comparative Analysis of China and the West" (Fundamentos culturales y éticos de la divergencia en la gobernanza de la IA: Un análisis comparativo de China y Occidente), *Política Internacional* VII, no. 1 (enero-marzo, 2025): 215-233. Publicado por el Center for American Studies, Fudan University, Shanghái. Recuperado de: https://cas.fudan.edu.cn/info/1212/21153.htm?utm_source=chatgpt.com

de estas tecnologías disruptivas. Esta lógica estatista, que contrasta nítidamente con el enfoque liberal predominante en Occidente, se fundamenta en la convicción de que solo una autoridad central fuerte puede prevenir los riesgos y las desviaciones que una evolución descontrolada de la IA podría entrañar para la armonía social y la "prosperidad socialista".

Este "confucianismo digital" permea las principales leyes e iniciativas regulatorias que configuran el marco normativo de la IA en China. Desde la Ley de Seguridad de Datos, que establece un régimen de clasificación de la información según su relevancia para la seguridad nacional y el desarrollo socioeconómico, hasta la Ley de Protección de la Información Personal, que si bien introduce salvaguardas para la privacidad individual, las supedita siempre a las exigencias superiores de la estabilidad social y el interés público, pasando por el proyecto de Ley de Inteligencia Artificial, que instituye un sistema de supervisión estatal omnímodo sobre el ciclo de vida de estas tecnologías, todos estos instrumentos jurídicos reflejan una filosofía regulatoria que prioriza el control político y la planificación centralizada por encima de la autonomía de los actores privados y la protección de los derechos individuales.

Las prioridades de gobernanza de China en materia de IA, centradas en el avance tecnológico, la promoción de aplicaciones prácticas y el fortalecimiento de la competitividad, encarnan los valores colectivistas del confucianismo. El énfasis en el progreso tecnológico al servicio del bienestar social, la priorización de las aplicaciones que mejoran la eficiencia económica y la calidad de vida, y el impulso a la competitividad nacional como medio para fortalecer el poder del Estado, reflejan una concepción de la IA como herramienta para alcanzar el ideal confuciano de un "mundo armonioso", donde el gobierno y las élites sociales colaboran en pos del bien común.

Esta visión se materializa en los métodos de gobernanza que caracterizan el enfoque chino. El centralismo gubernamental, manifestado en el rol directivo del Partido-Estado en la formulación de políticas, la asignación de recursos y la regulación de la IA, evoca la noción confuciana del gobierno como "Estrella Polar" que guía el devenir social²⁶⁵. El funcionalismo y la orientación al desempeño, plasmados en el énfasis en las aplicaciones prácticas y los resultados tangibles, reflejan la prioridad que el confucianismo otorga a las contribuciones al bienestar colectivo. Y el

²⁶⁵ Ibid, 222-226.

énfasis en los intereses sociales sobre los individuales, evidenciado en la subordinación de la privacidad a la seguridad y el interés público, encarna el ideal confuciano de armonía entre individuo y sociedad²⁶⁶.

En última instancia, el "confucionismo digital" que impregna la aproximación china a la regulación de la IA constituye un desafío y un contrapunto ineludible para el orden jurídico internacional. Su énfasis en la centralidad del Estado, la armonía social y la subordinación de los derechos individuales al interés colectivo interpela los fundamentos mismos del constitucionalismo liberal y plantea interrogantes de calado sobre la gobernanza global de unas tecnologías cuyo impacto trasciende las fronteras nacionales.

E.- Tensiones y Desafíos

El análisis de la lógica centralizada que subyace a la gobernanza china de la inteligencia artificial, moldeada por lo que algunos han denominado "confucionismo digital", nos invita a una reflexión profunda sobre las tensiones y los desafíos que este enfoque plantea frente al paradigma liberal predominante en Occidente. Lejos de caer en simplificaciones maniqueas o juicios de valor apresurados, este apartado busca ofrecer una mirada matizada y equilibrada sobre las diferencias fundamentales entre ambos modelos, sus potenciales ventajas y riesgos, además de las implicaciones para la construcción de un marco regulatorio global para la IA.

En el núcleo de estas tensiones, se sitúan las divergencias filosóficas y jurídicas que separan al confucionismo del liberalismo. Mientras que el primero prioriza la armonía social, la centralidad del Estado y la subordinación de los derechos individuales al interés colectivo, el segundo se erige sobre los pilares de la autonomía individual, la limitación del poder estatal y la protección de las libertades fundamentales. Estas diferencias, enraizadas en tradiciones culturales y trayectorias históricas diversas, se proyectan inevitablemente sobre los enfoques regulatorios de la IA, dando lugar a modelos que, si bien persiguen objetivos comunes como el progreso tecnológico y el bienestar social, difieren en sus métodos y prioridades.

²⁶⁶ Ibid.

El modelo chino, caracterizado por el centralismo gubernamental, el funcionalismo y la orientación al desempeño, presenta ciertas ventajas en términos de eficiencia, coordinación y coherencia. La capacidad del Estado para movilizar ingentes recursos, alinear a los actores clave y mantener una supervisión estrecha sobre el ecosistema de la IA augura una trayectoria de innovación acelerada y un despliegue masivo de estas tecnologías en sectores estratégicos. Asimismo, la subordinación de los intereses particulares al bien común y la priorización de los resultados tangibles sobre las consideraciones individuales pueden facilitar la implementación de políticas ambiciosas y la consecución de objetivos a gran escala.

Sin embargo, estas potenciales ventajas no están exentas de riesgos y desafíos desde la óptica liberal. La concentración del poder decisorio en el Estado y el Partido, la opacidad de los procesos de gobernanza y la ausencia de contrapesos institucionales suscitan interrogantes sobre la rendición de cuentas y la protección frente a posibles abusos. La primacía de los intereses colectivos sobre los derechos individuales plantea dilemas sobre la salvaguarda de la autonomía y la privacidad en un contexto de creciente datificación y vigilancia tecnológica. Y el énfasis en la eficiencia y el desempeño puede relegar a un segundo plano consideraciones éticas y sociales que, desde una perspectiva liberal, son indisociables del desarrollo responsable de la IA.

Otro desafío radica en la sostenibilidad económica y la rivalidad geopolítica con potencias occidentales. Los vetos de EE. UU. a la venta de semiconductores avanzados²⁶⁷ y tecnologías de punta a China podrían ralentizar la escalada de su IA, generando una bifurcación de cadenas de suministro²⁶⁸ y normas internacionales divergentes.

F.- Proyecciones: Consolidación de un Modelo Centralista y Reacciones Globales

Todo indica que el régimen normativo chino sobre IA no dará marcha atrás en su centralización y control. Antes bien, cabe prever una profundización de la “gestión integral” de los datos, la sofisticación de las “Smart Courts” y la integración de tecnologías de aprendizaje

²⁶⁷ William Alan Reinsch, Matthew Schleich, y Thibault Denamiel, "Insight into the U.S. Semiconductor Export Controls Update," Center for Strategic and International Studies (CSIS), 20 de octubre de 2023, <https://www.csis.org/analysis/insight-us-semiconductor-export-controls-update>

²⁶⁸ Terry E. Chan et al., "Technology and Geopolitics: What If The Semiconductor Industry Bifurcates?" S&P Global Ratings, 14 de noviembre de 2022, <https://www.spglobal.com/ratings/en/research/articles/221114-technology-and-geopolitics-what-if-the-semiconductor-industry-bifurcates-12557030>

automático en diferentes escalones administrativos. El objetivo declarado: robustecer el tejido industrial, reducir la dependencia de proveedores extranjeros y garantizar la “cohesión social” a través de la tecnovigilancia y posicionarse como la primera potencia en inteligencia artificial para el 2030.

A su vez, las potencias occidentales, incluida la Unión Europea, observan con recelo esta consolidación de la IA autoritaria. No solo se plantean tensiones comerciales y estándares incompatibles de gobernanza, sino que, también, emergen dilemas éticos sobre la legitimidad de intercambiar datos o adoptar soluciones “made in China” en escenarios sensibles.

En definitiva, el marco normativo chino combina leyes de protección de datos aparentemente progresistas con amplias facultades estatales que relativizan la privacidad y la independencia judicial. Esta singular fusión estatista, unida a la ambición geopolítica, define un modelo hegemónico de la IA que rivaliza con el liberalismo estadounidense y el garantismo europeo, con profundas implicaciones para la configuración del futuro orden digital global.

2.8.3. Análisis Comparativo: Estados Unidos versus China en la Regulación de la IA

La comparativa entre el **enfoque liberal** de Estados Unidos y el **modelo centralista** de la República Popular China en materia de inteligencia artificial (IA) encierra un abanico de disimilitudes filosóficas, institucionales y normativas de enorme calado. Ambas jurisdicciones, protagonistas indiscutibles en el terreno tecnológico global, comparten la determinación de erigirse en potencias hegemónicas en el desarrollo y la aplicación de la IA. Sin embargo, la forma de articular las reglas de juego para que esa transformación tecnológica no colisione abiertamente con la seguridad, la economía o los valores sociales resulta diametralmente distinta.

A. Fundamentos Filosóficos

➤ Tradición Liberal frente a Centralismo Estatal

- **Estados Unidos:** se adhiere históricamente a una **posición liberal y pro-mercado**, promoviendo la autorregulación y la intervención mínima del Estado en asuntos de innovación. Su matriz conceptual confía en que la competencia empresarial impulsará

la eficacia y la competitividad de la IA. Aunque se constatan algunas directrices sectoriales, la ausencia de una ley federal unificada refleja la convicción de que un marco fragmentado, pero menos rígido, favorece la adaptabilidad al vertiginoso cambio tecnológico,

- **China:** emerge un **modelo centralizado** donde el **Partido Comunista** lidera y orienta la estrategia nacional de IA, supeditándola al bienestar social y la estabilidad política. El **Estado** diseña planes de alto nivel —v. gr. el *New Generation Artificial Intelligence Development Plan*— y coordina a las corporaciones locales para encauzar la innovación en consonancia con las metas colectiva. El control central del partido se traduce en regulaciones detalladas, amparadas en la idea de que la IA debe servir de herramienta para el desarrollo nacional y la gobernanza social.

➤ Apuesta por la Innovación vs. Enfoque de Seguridad

- **EE. UU.** resalta el **potencial económico** de la IA, considerándola clave para mantener el liderazgo global. Aunque la doctrina hace referencia a la “IA responsable” y la “protección de datos”, el motor principal es la competitividad tecnológica, no la formulación de un corpus uniforme de derechos,
- **China** prioriza la **seguridad estatal** y la cohesión social, junto a la supremacía global en IA. Como parte de esa estrategia, la vigilancia y el acceso a ingentes volúmenes de datos son contemplados como garantías de un control eficiente, legitimados por la primacía de protección del interés colectivo.

B. Estructura Institucional y Niveles de Regulación

➤ El Federalismo Estadounidense y la Fragmentación

En **Estados Unidos**, el poder regulatorio en IA se reparte de forma **multicéntrica**:

1. **Agencias federales** (FDA, DOT, CFPB) emiten guías específicas para la IA en cada sector.
2. **Los estados** dictan leyes parciales (p. ej., la CCPA en California), generando un mosaico de normas sobre privacidad y uso de datos.

3. **El Congreso y la Casa Blanca** publican lineamientos no vinculantes —como el *Blueprint for an AI Bill of Rights*— para encauzar la autorregulación.

Este entramado **disperso** da cuenta de la primacía de la lógica liberal, confiando en que los diversos órganos y la litigación en cortes provean correcciones puntuales donde se vean conculcados los derechos fundamentales.

➤ **El Poder Central en la RPCh y la Jerarquía Vertical**

China opera con un **alto grado de centralización**:

1. El **Consejo de Estado** y las **comisiones del Partido** definen la estrategia global, fijando metas y plazos para la adopción de la IA.
2. Las **Leyes nacionales** (Data Security Law, PIPL) se aplican uniformemente en todo el territorio, dejando escaso margen a las regiones.
3. Organismos como la Administración del Ciberespacio de China (CAC) emiten reglamentos pormenorizados —por ejemplo, la gestión de la síntesis profunda— con supervisión directa del aparato gubernamental.

En contraposición al federalismo estadounidense, **Beijing** mantiene una posición dominante, simplificando la adopción de políticas y asegurando una coherencia central, a costa de la diversidad local y la independencia de las instituciones.

C. Impacto Geopolítico y Colisión de Estándares

➤ **Exportación de Modelos IA y Pugna por la Hegemonía Digital**

La dicotomía entre un **ecosistema liberal** (EE. UU.) y otro **centralista** (China) se traslada a la escena mundial:

- Las big tech estadounidenses (Google, Microsoft, Amazon) ofrecen soluciones que se rigen por el criterio de “compliance selectiva” con diversas jurisdicciones,

- Las empresas chinas, bajo fuerte tutela gubernamental, expanden su influencia en países en desarrollo (África, Sudeste Asiático, etc.), **exportando** tecnologías de reconocimiento facial, telecomunicaciones, y análisis masivo que se integran a esos sistemas nacionales.

Ello genera una fractura potencial en la que las normas y los estándares practicados por Estados Unidos y China pueden entrar en contradicción. En efecto, si un país adopta soluciones tecnológicas chinas, podría verse constreñido a permitir la transferencia de datos a servidores bajo supervisión de Pekín.

D. Conclusión: Dos Polos Opuestos que Definen el Rumbo de la IA Global

El **contraste** entre el liberalismo estadounidense —caracterizado por una fragmentación regulatoria y la preferencia por el mercado— y el modelo chino **estatista y centralizado** no podía ser más marcado. Las implicaciones prácticas se extienden desde la protección de la privacidad y los derechos del imputado en contextos penales, hasta la supervisión de las aplicaciones de IA en la función judicial.

- **En Estados Unidos**, la ausencia de un mando unificado produce un juego cambiante de iniciativas, con predominio del caso a caso y la litigación *ex post*. Se prioriza la innovación, quedando la contención de riesgos en manos de agencias sectoriales y eventuales demandas civiles,
- **En China**, el peso del Estado y la ideología del partido condicionan toda la trama legal. Lejos de la mera recomendación, el plan de IA se convierte en un mandato vertical, apoyado en leyes que facultan la vigilancia e intervención pública.

A nivel **geopolítico**, esta disparidad configura un escenario donde cada polo pretende imponer su impronta, generando colisiones de estándares. Los países que se integran a la órbita de Washington o de Pekín asumen concepciones distintas de cómo regular la IA. En medio, la Unión Europea ha propuesto un enfoque basado en la dignidad y la no discriminación, con vocación de irradiar su influencia más allá de las fronteras europeas.

El **equilibrio final** —si uno se inclina por la competencia libre y la autorregulación, o por el control estricto y centralizado— dependerá, en gran medida, del rumbo que tomen las disputas

comerciales y las alianzas estratégicas en los próximos años. Sea como fuera, el análisis comparativo muestra que **Estados Unidos y China** definen dos **paradigmas antagónicos** de la **regulación de la IA**, influidos por raíces históricas y políticas que trascienden el mero ámbito tecnológico.

2.8.4. Conclusiones Finales: Convergencias y Divergencias respecto del Paradigma de la Unión Europea

El análisis comparativo de los modelos regulatorios de Estados Unidos y China en el ámbito de la Inteligencia Artificial (IA) pone de manifiesto la existencia de **dos grandes polos** que, desde raíces históricas y culturales disímiles, se esfuerzan por moldear el futuro de la IA a escala global. Ahora bien, **la Unión Europea (UE)** emerge, a su vez, con un **enfoque propio**, articulado en el **Reglamento Europeo sobre la IA** (EU AI Act) y otros instrumentos normativos, donde la centralidad de los **derechos fundamentales** (dignidad humana, privacidad, no discriminación) adquiere un relieve prioritario. A renglón seguido, se examinan las **convergencias y divergencias** de ambos modelos con el de la UE, extrayendo implicaciones y delineando posibles escenarios de interacción.

A.- Convergencias Potenciales

- **Reconocimiento de la Relevancia de la IA:** tanto en Estados Unidos como en China, se admite sin ambages la trascendencia de la IA como vector de competitividad y como fuerza transformadora de la economía. En la UE, el Libro Blanco sobre IA de 2020 reconoce la necesidad de incentivar la innovación y fortalecer el tejido industrial. Así, las tres superpotencias comparten el objetivo de dominar la IA para afianzar su prosperidad y liderazgo global.
- **Búsqueda de un Certo Esquema de Supervisión:** en EE. UU., a pesar del liberalismo imperante, iniciativas como el *Blueprint for an AI Bill of Rights* expresan la voluntad de mitigar sesgos y asegurar la seguridad de los sistemas. **China**, por su parte, obliga a los proveedores a observar normas que eviten el uso indebido de la IA (p. ej., provisiones sobre la síntesis profunda). Y de su lado, la **UE** enfatiza la clasificación de riesgo y la vigilancia humana en aplicaciones de IA de alto riesgo. En consecuencia, puede al menos

vislumbrarse una convergencia parcial en torno a la idea de que la IA, por su potencia disruptiva, requiere cierto grado de control legal o reglamentario.

3. **Afirmación de la Importancia del Sector Privado:** pese a sus profundas diferencias políticas, las tres jurisdicciones (UE, EE. UU., China) reconocen que la colaboración entre gobierno y empresas privadas es esencial. La UE recurre a la inversión público-privada e impulsa *sandboxes* regulatorios, los Estados Unidos confían en la autorregulación y la competitividad y China moviliza a conglomerados como Tencent o Alibaba, canalizando sus recursos hacia la “misión” colectiva.

B. Divergencias Fundamentales

➤ Acento en Derechos Fundamentales vs. Otros Objetivos

La divergencia fundamental entre la Unión Europea, Estados Unidos y China en materia de regulación de la inteligencia artificial reside, en esencia, en la primacía otorgada a los derechos fundamentales frente a otros objetivos de política pública. **Mientras que el AI Act europeo erige la protección de las libertades individuales como su piedra angular, el modelo estadounidense prioriza la innovación y la competitividad y el enfoque chino antepone el control estatal y la estabilidad social.** Esta discrepancia, lejos de ser meramente técnica, hunde sus raíces en contrastes filosóficos y políticos más profundos que moldean la visión de cada jurisdicción sobre el papel de la tecnología en la sociedad.

En el corazón del paradigma europeo late una concepción kantiana de la dignidad humana como fin en sí mismo, nunca como mero medio. El AI Act, al consagrarse a principios como la no discriminación, la privacidad, la seguridad jurídica y la supervisión humana efectiva, busca salvaguardar esa dignidad frente a los riesgos de una IA descontrolada. La prohibición de sistemas que empleen técnicas subliminales, exploten vulnerabilidades o evalúen la fiabilidad a partir de datos biométricos revela una voluntad de trazar líneas rojas infranqueables, más allá de las cuales ninguna ganancia en eficiencia justificaría el sacrificio de derechos fundamentales. Este "personalismo tecnológico", si se permite la expresión, aspira a humanizar la revolución digital, sometiéndola al imperio de los valores constitucionales.

En el polo opuesto, Estados Unidos abraza una visión utilitarista donde la innovación se erige como valor supremo. La fragmentación regulatoria, la preferencia por la autorregulación empresarial y la intervención *ex post* de las agencias y tribunales reflejan una fe casi religiosa en la capacidad del mercado para corregir sus propias disfunciones. Desde esta óptica, el progreso tecnológico es un fin en sí mismo y cualquier intento de encorsetarlo preventivamente con reglas rígidas se percibe como una amenaza a la competitividad y el liderazgo global. La protección de derechos se concibe, pues, como una consideración *a posteriori*, a invocar cuando se haya consumado una vulneración flagrante, pero no como un principio rector que deba modelar de antemano la arquitectura de los sistemas de IA.

China, por su parte, encarna un confucionismo digital donde el Partido-Estado se erige en árbitro supremo de la bondad y legitimidad de toda innovación. El New Generation Artificial Intelligence Development Plan, con su énfasis en la IA como herramienta de gobernanza social y su subordinación de los derechos individuales al interés colectivo, evoca ecos de un neoconfucianismo adaptado a la era de los algoritmos. En este esquema, la autonomía personal se difumina frente al imperativo de la armonía y la cohesión social.

Estas divergencias, más que un simple desencuentro regulatorio, reflejan cosmovisiones antagónicas sobre la relación entre tecnología, individuo y sociedad. ¿Debe la IA servir prioritariamente a la dignidad humana, a la innovación disruptiva o al control social? La respuesta a esta pregunta condiciona irremediablemente el enfoque normativo adoptado.

En un mundo globalizado, sin embargo, estos modelos no pueden coexistir en compartimentos estancos. El carácter transfronterizo de la IA y su impacto en ámbitos como el comercio internacional, los derechos humanos o la seguridad obligan a buscar un mínimo común denominador. La Unión Europea, con su "efecto Bruselas", aspira a exportar sus estándares garantistas más allá de sus fronteras, generando una convergencia normativa por la vía del mercado interior. Estados Unidos, celoso de su primacía tecnológica, se resiste a cualquier alineamiento que pueda mermar su ventaja competitiva. Y China, en su afán por moldear un orden digital sinocéntrico, promueve activamente la adopción de sus soluciones de IA en países en desarrollo.

Este juego de equilibrios inestables augura un futuro donde la gobernanza global de la IA se dirimirá no solo en los parlamentos y los tribunales, sino, también, en las mesas de negociación comercial y en los foros geopolíticos. El reto, en última instancia, radica en alumbrar un marco

regulatorio que, sin ahogar la innovación ni sucumbir a la tentación del control absoluto, preserve esos derechos fundamentales que nos definen como civilización.

➤ **Privacidad y Protección de Datos Personales**

La Unión Europea, heredera de una tradición humanista que hunde sus raíces en la Ilustración y la Declaración Universal de los Derechos Humanos, ha erigido en torno a la autodeterminación informativa un baluarte inexpugnable. El Reglamento General de Protección de Datos (RGPD), con su énfasis en la transparencia, la finalidad y minimización, encarna una concepción quasi-proprietaria de la información personal como emanación de la dignidad ontológica del individuo. El AI Act, al someter los sistemas de alto riesgo a exigencias reforzadas de gobierno y calidad de los datos, no hace sino ahondar en este compromiso con la sacralidad de la esfera íntima.

Estados Unidos, por su parte, ofrece un caleidoscopio de enfoques fragmentarios donde la protección de la privacidad oscila según el sector y la coyuntura. La ausencia de un marco federal omnicomprendioso al estilo europeo deja la tutela de los datos personales a merced de un mosaico de leyes estatales, guías sectoriales y compromisos voluntarios de la industria. Esta dispersión normativa, aunque mitigada en parte por hitos como la California Consumer Privacy Act, refleja una concepción de la privacidad más cercana a la de un bien negociable que a la de un derecho fundamental inderogable. La primacía de la libertad individual y la desconfianza hacia la intervención estatal, pilares del ethos estadounidense, permean también el ámbito de la autodeterminación informativa.

La República Popular China, en cambio, aborda la cuestión desde una óptica colectivista enraizada en su acervo filosófico y su peculiar trayectoria histórica. En el epicentro de su modelo regulatorio se sitúa la noción de que los datos, más que un atributo individual, constituyen un activo estratégico para el desarrollo nacional y la gobernanza social. La Ley de Seguridad de Datos y la Ley de Protección de la Información Personal, aunque introducen salvaguardas frente a injerencias arbitrarias, consagran la prerrogativa estatal para acceder y explotar la información en aras del bien común. Esta asimetría entre las garantías del ciudadano y las potestades del Leviatán digital responde a una cosmovisión donde el interés colectivo se erige en valor supremo, relegando la autonomía individual a un papel subsidiario.

➤ Transparencia y Rendición de Cuentas

No menos profundas son las divergencias en materia de transparencia y rendición de cuentas de los sistemas de IA. La Unión Europea, consciente de la opacidad consustancial a los algoritmos complejos, ha convertido la explicabilidad en uno de los pilares de su estrategia regulatoria. El *AI Act* impone a los proveedores de sistemas de alto riesgo la obligación de facilitar información clara y adecuada sobre su funcionamiento, así como de implementar medidas de trazabilidad y registro de eventos. Se aspira, en suma, a conjurar la amenaza de una "caja negra" ingobernable, sometiendo la IA al escrutinio público y al imperio de la ley.

En contraste, tanto el modelo estadounidense, como el chino, se muestran más refractarios a destapar la "caja de Pandora" algorítmica. En Estados Unidos, la renuencia a imponer obligaciones de transparencia excesivamente onerosas parte de la premisa de que la opacidad es, en cierta medida, el precio a pagar por la innovación. Los secretos comerciales y la competitividad se esgrimen como justificación para preservar la confidencialidad de unos sistemas cuya eficacia depende, precisamente, de su inescrutabilidad.

China, por su parte, aborda la transparencia desde un prisma eminentemente estatista. Aunque leyes como la de Seguridad de Datos imponen obligaciones de gestión de riesgos e información al usuario, el grueso de la explicabilidad se concibe como una prerrogativa de las autoridades, no como un derecho ciudadano. Son los organismos públicos quienes pueden exigir acceso a la lógica algorítmica, no en aras de la rendición de cuentas democrática, sino de la preservación de la estabilidad y los valores socialistas. La "caja negra", desde esta óptica, se vuelve transparente para el ojo avizor del partido, pero permanece opaca para el escrutinio popular.

➤ Intervención *Ex Ante* y Control *Ex Post*

La tensión entre intervención *ex ante* y control *ex post* es otro punto de divergencia capital. La Unión Europea, fiel a su vocación garantista, apuesta por una regulación preventiva y pormenorizada que minimice los riesgos antes de que se materialicen. El *AI Act* establece un proceso de evaluación de conformidad previo a la comercialización de los sistemas de alto riesgo, así como requisitos estrictos de gestión de riesgos, gobierno de datos y documentación técnica. Se

trata, en esencia, de inocular principios éticos en el diseño mismo de la IA, modulando su desarrollo para embridar su potencial disruptivo.

Estados Unidos, en cambio, fía la supervisión de la IA a una combinación de autorregulación empresarial y vigilancia reactiva de las agencias sectoriales. Se confía en que el propio mercado, alentado por el temor al riesgo reputacional y la amenaza de litigios, acabará decantándose por sistemas seguros y fiables. La intervención pública se reserva para los casos más flagrantes, siguiendo un enfoque de "palo y zanahoria" donde la innovación se recompensa y los abusos se sancionan *a posteriori*.

China, pese a su apariencia de control férreo, también se inclina por un modelo más reactivo que preventivo. Si bien leyes como la de Seguridad de Datos imponen obligaciones *ex ante*, como la categorización de la información y la evaluación de riesgos, el grueso del aparato regulatorio se orienta a la supervisión *ex post* y la respuesta rápida ante desviaciones. La Administración del Ciberespacio de China (CAC), con sus amplios poderes de inspección y sanción, encarna esa filosofía de "rienda larga" que permite un amplio margen de experimentación, siempre bajo la atenta mirada del Estado.

➤ Fomento de la Innovación

Por último, la discrepancia en cuanto al fomento de la innovación y asunción de riesgos revela contrastes fundamentales en la concepción del progreso tecnológico. Estados Unidos, imbuido del mito del excepcionalismo y la destrucción creativa schumpeteriana, abraza la disruptión como un fin en sí mismo. La minimización de las trabas regulatorias, unida a un ecosistema de capital riesgo y un marco de propiedad intelectual robusto, propicia una cultura emprendedora donde el fracaso no es un estigma, sino un rito de iniciación. Este caldo de cultivo ha catapultado a Silicon Valley a la vanguardia mundial de la IA, aun a costa de una cierta desatención a sus externalidades sociales.

La Unión Europea, en cambio, pugna por conciliar innovación y precaución en un equilibrio inestable. Consciente de su rezago en la carrera por la supremacía tecnológica, Bruselas

ha lanzado iniciativas como el Plan Coordinado sobre la Inteligencia Artificial²⁶⁹ para estimular la investigación y el desarrollo de una IA puntera. Al mismo tiempo, el énfasis en la gestión de riesgos, el principio de precaución y la protección de los derechos fundamentales actúa como contrapeso a los excesos de una disruptión descontrolada. Se aspira, en suma, a un modelo de innovación "responsable" donde los avances se ponderen a la luz de sus implicaciones éticas y sociales.

China, por su parte, concibe la innovación como una herramienta de empoderamiento nacional, subordinada siempre a los imperativos del desarrollo y la estabilidad. A través de planes quinquenales y directrices sectoriales, el Partido-Estado fija las prioridades estratégicas y canaliza los recursos hacia las áreas de la IA consideradas vitales para la proyección geopolítica del país. Se fomenta la asunción de riesgos, sí, pero dentro de los cauces marcados por el interés colectivo y bajo la tutela de un Estado cuya misión autoasignada es la de timonel del progreso. El resultado es un ecosistema de innovación dirigida, donde la creatividad se encarrila hacia metas socialistas y la iniciativa privada se funde con el designio público en una simbiosis sin parangón.

En última instancia, estas divergencias, más que un mero desencuentro técnico, reflejan cosmovisiones antagónicas sobre el individuo, la sociedad y el porvenir. La Unión Europea, heredera de la tradición liberal-personalista, sitúa la dignidad humana en el epicentro de su proyecto regulatorio. Estados Unidos, adalid del libre mercado, fía el rumbo de la IA a las fuerzas de la innovación y la competencia. Y China, abanderada de un neo-colectivismo tecnológico, la concibe como un instrumento de gobernanza social y engrandecimiento nacional.

Armonizar estos modelos en un corpus normativo global coherente se antoja una quimera. Quizá el desafío radique, no tanto en homogeneizar lo que es diverso por naturaleza, sino en alumbrar un marco de convivencia que, desde el respeto a la pluralidad axiológica, permita aprovechar el potencial transformador de la IA sin socavar los cimientos de nuestra humanidad compartida. Un reto mayúsculo que interpela a juristas, filósofos, tecnólogos y ciudadanos por

²⁶⁹ Comisión Europea, *Revisión de 2021 del plan coordinado sobre la inteligencia artificial: Comunicación de la Comisión al Parlamento Europeo, al Consejo Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones*, COM(2021) 205 final, Bruselas, 21 de abril de 2021. Recuperado de: <https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review>

igual, pues lo que está en juego no es solo el gobierno de una revolución, sino el porvenir mismo de nuestra civilización en la era algorítmica.

C. Epílogo

El análisis comparativo revela que el **enfoque europeo** conforma un **tercer sendero** entre las dos hegemónias, abogando por la centralidad de la persona y la sujeción de la IA a normas innegociables de **transparencia, supervisión humana y protección de datos**. Mientras el **modelo estadounidense** podría, por convergencia parcial, evolucionar hacia estándares más próximos a los de la UE, el **modelo chino** se aleja en lo esencial, al mantener la supremacía del interés del Partido-Estado. Así, el **AI Act** de la Unión Europea se perfila como un contrapeso regulatorio capaz de inspirar a otras regiones o de generar colisiones con las potencias que optan por la autorregulación o la instrumentalización estatal de la IA.

En definitiva, la “**convergencia o divergencia**” con la Unión Europea depende de la sintonía que cada potencia esté dispuesta a admitir en materia de derechos y libertades. De momento, la divergencia domina el panorama: Estados Unidos y China contemplan la IA desde prismas que difieren sustancialmente del garantismo. No obstante, los retos comunes —sesgos, ciberseguridad, responsabilidad por daños— pueden forzar, a medio plazo, cierta confluencia. La pieza clave será, sin duda, la presión internacional y la aceptación o no de la premisa fundacional europea: que la IA debe, ante todo, **servir al ser humano** y respetar sus derechos más elementales.

2.9. Desafíos y Perspectivas Futuras de la Regulación de la IA en la Justicia Europea

La irrupción de la Inteligencia Artificial (IA) en la esfera judicial se ha consolidado como uno de los desafíos más ambiciosos y fascinantes que enfrentan las democracias constitucionales en el presente siglo. Aunada a la complejidad propia de la administración de justicia—tradicional garante de los derechos y libertades—la IA introduce nuevas dinámicas de automatización que, a un mismo tiempo, prometen mayor eficiencia y plantean riesgos inéditos para la autonomía decisoria, la equidad procesal y la dignidad de la persona. Esta tensión dibuja un escenario en el que la **regulación** deviene indispensable, pero —según algunos críticos— puede erigirse en un

freno a la innovación acelerada que hoy caracteriza el campo de la IA. De ahí la relevancia de una “**tercera vía garantista**” que la Unión Europea (UE) ha venido defendiendo: un paradigma que anhela armonizar el propósito de la innovación tecnológica con la salvaguarda de los principios constitucionales que vertebran el Estado de Derecho.

2.9.1. Paradojas de la IA en la Justicia: Promesa y Cautela

La IA avanza con ímpetu en el ámbito judicial. Las soluciones de “justicia predictiva” o de análisis automatizado de casos se han multiplicado, apuntando a una reducción de la carga de trabajo de los tribunales, al aumento de la coherencia jurisprudencial y a la optimización de los tiempos de respuesta. Sin embargo, **la adopción de sistemas algorítmicos en un escenario tan delicado—donde se decide sobre derechos fundamentales—conlleva riesgos sustanciales**. El principal radica en la posible **erosión de garantías** como la independencia del juez, la imparcialidad y la tutela judicial efectiva, si se entroniza la lógica de la “**caja negra**”.

A esta realidad subyace una paradoja fundamental: **la IA, capaz de propulsar la efectividad judicial, requiere un andamiaje regulatorio robusto que impida desviaciones antidemocráticas**. Pero, en la medida en que la UE articula requisitos estrictos de transparencia, supervisión humana y calidad de datos, aumenta la carga para los desarrolladores, lo que podría restar competitividad frente a jurisdicciones con normas más laxas²⁷⁰. Algunos sostienen que este “freno” regulatorio, aunque moralmente loable, dificulta la experimentación y aparta la inversión a entornos más permisivos. Tal tensión revela que no basta con generalidades éticas: la UE debe **equilibrar** sus altos estándares garantistas con incentivos que no desanimen la innovación.

2.9.2. La “Tercera Vía Garantista”: Fundamentos y Dilemas

La concepción de una “**tercera vía garantista**” en la regulación de la IA aplicada a la justicia hunde sus raíces en la tradición humanista y constitucionalista de la Unión Europea (UE). Se trata de un camino intermedio entre:

²⁷⁰ Richard Oubělický, "Europe and AI: Causes and Implications of Europe Losing Ground in the Race for AI" Security Outline, 13 de marzo de 2024, <https://www.securityoutlines.cz/europe-and-ai-part-i/>

1. El **liberalismo autorregulador** de Estados Unidos, que prefiere controles ex post y confía en la litigación y la competencia de mercado para corregir abusos.
2. El **centralismo confuciano** de China, que instrumentaliza la IA bajo la égida del Estado y prioriza la estabilidad social por encima de la esfera individual.

Frente a esos dos extremos, la UE impulsa un modelo donde la innovación tecnológica se somete a controles preventivos, de manera que los algoritmos **no lesionen** derechos fundamentales ni vacíen de contenido la independencia judicial. Sin embargo, esta concepción—por loable que sea—afronta sus propios dilemas y tensiones, que a continuación se exponen detalladamente.

A. Bases Filosófico-Jurídicas: el Estado de Derecho como Escudo frente a la Automatización Deshumanizadora

El núcleo de esta tercera vía bebe de los valores constitucionales consagrados en la Carta de los Derechos Fundamentales de la UE, en particular:

- **Dignidad humana** (art. 1 CDFUE): ninguna innovación tecnológica puede relegar a la persona a la condición de objeto pasivo de decisiones ininteligibles. De ahí que la UE exija transparencia y explicabilidad en la IA judicial,
- **Derecho a la tutela judicial efectiva** (art. 47 CDFUE): la intervención del juez no se puede suplantar por un algoritmo que decida sin criterio humano. Incluso cuando un sistema predictivo aporte un “score” o “recomendación”, debe haber un magistrado que ejercite su prudencia y evalúe el contexto específico,
- **Igualdad y no discriminación** (art. 21 CDFUE): si la IA se basa en datos históricos cargados de sesgos, cabe el riesgo de reproducir o intensificar prejuicios contra minorías o grupos vulnerables. La UE prohíbe, por tanto, prácticas algorítmicas que discriminen o perpetúen desigualdades (Reglamento Europeo sobre la IA, art. 10)

Esta impronta garantista, derivada de un acervo iusfilosófico que combina el **personalismo** (cada persona como fin en sí mismo) con la **responsabilidad pública** ante las transformaciones tecnológicas, justifica la imposición de **líneas rojas** a la IA en la justicia.

B. Tensión entre la Salvaguarda de Derechos y la Competitividad Tecnológica

El primer gran dilema brota de la sospecha de que una regulación *ex ante* estricta—pilar de la vía garantista—**podría ralentizar la innovación**. Mientras Estados Unidos deja gran espacio a la autorregulación y la corrección judicial *a posteriori* (el caso *COMPAS* ilustra su pragmatismo), la UE ha erigido un entramado exigente de obligaciones:

- **Clasificación por niveles de riesgo** (arts. 6 y 7 de la Ley de IA)
- **Supervisión humana y trazabilidad** (arts. 13 y 14)
- **Gestión rigurosa de datos** (art. 10)

Estos requerimientos pueden implicar costes adicionales para desarrolladores, que podrían buscar sedes en otras jurisdicciones más laxas. Surge, por ende, la siguiente cuestión: **¿Estará dispuesta la UE a asumir la pérdida de velocidad en la carrera global de la IA a cambio de blindar sus principios constitucionales?** Sus instituciones parecen creer que sí, argumentando que la legitimidad y la confianza a largo plazo valen más que los réditos inmediatos.

C. Miedo al Exceso de Cautelas: ¿Riesgo de un “Formalismo Paralizante”?

Otro dilema afecta al posible **exceso de formalismos** que, en aras de evitar cualquier conculcación de derechos, terminen por burocratizar la innovación. Algunos críticos subrayan que la justicia europea, ya saturada, podría volverse todavía más lenta si cada nueva herramienta algorítmica necesita pasar por una densa evaluación de conformidad y por un mecanismo de registro (arts. 43 y 49 de la Ley de IA). ¿No agravaría esto la congestión judicial, en vez de aligerarla?

Quienes defienden la vía garantista replican que **el Estado de Derecho no se presta a atajos**: mejor un tránsito lento, pero seguro, que el “deslumbre” de soluciones rápidas con resultados opacos y discriminatorios. Al fin y al cabo, la justicia no es una fábrica de sentencias; es el pilar sobre el que descansa la legitimidad del ordenamiento jurídico.

D. ¿Dónde Situar el Listón de la Intervención Humana?

La tercera vía garantista insiste en la **supervisión humana** como núcleo del uso de IA en la justicia (art. 14 de la Ley de IA). Empero, un dilema subyace: **¿hasta qué punto el juez debe (o puede) fiarse de la recomendación algorítmica en asuntos complejos?**

- Si la supervisión es superficial, existe el sesgo de automatización: el magistrado, abrumado de trabajo, podría aceptar acríticamente la salida del algoritmo,
- Si la supervisión es meticulosa y desconfía sistemáticamente del sistema, la IA se vuelve irrelevante o entorpece la rutina judicial.

Por ende, la UE necesita encontrar un **equilibrio**: formar a los jueces y dotarlos de herramientas de interpretabilidad que permitan un control razonable—no meramente simbólico—de la actuación algorítmica. El riesgo es que la mera exigencia legal de “supervisión humana” se convierta, en la práctica, en un formalismo vacío, si los jueces no reciben formación tecnológica profunda.

E. Legitimidad Democrática y Reputación Institucional

Un último aspecto esencial del que emana la “tercera vía garantista” es la **legitimidad democrática**. El proceso de adopción de normas sobre IA (el AI Act, el Libro Blanco, las directrices de la CEPEJ) se ha nutrido de consultas públicas, informes de expertos y debates en el Parlamento Europeo. Esa metodología encarna la transparencia y la participación ciudadana, algo muy distinto a los procesos decisorios de otros regímenes.

Así, la UE busca —mediante su cuidada elaboración normativa— **consolidar su reputación como adalid de la protección de derechos** en la era digital. Lo hace sabiendo que la tecnología de IA en la justicia tiene un impacto directo en la vida de los justiciables, por lo que la sociedad europea está vigilante ante cualquier tentación de “automatizar la justicia” de forma inescrutable.

2.9.3. El Temor a la “Justicia Codificada”: la Trampa de la Uniformización

Uno de los mayores recelos ante la IA judicial proviene del riesgo de caer en una “**justicia codificada**” que priorice la aplicación mecánica de patrones estadísticos sobre la singularidad del caso. La justicia no es —ni puede ser— una mera ecuación. Las razones son profundas:

1. **Discrecionalidad y ponderación:** el acto de juzgar va más allá de un silogismo lógico; entraña la ponderación de principios, la valoración de factores contextuales y la adaptación de la norma a la especificidad fáctica. Autores como Dworkin sostienen que aun en los llamados “casos fáciles” hay un margen axiológico y de equidad.
2. **Dignidad humana:** un juicio justo no se limita a la exactitud de la calificación penal o la cuantía indemnizatoria. Implica brindar al justiciable la oportunidad de ser oído y valorado en su irrepetible condición personal (art. 6 CEDH).
3. **Responsabilidad y legitimidad:** cuando la IA se concibe como un oráculo, el juez podría caer en un sesgo de automatización, abdicando de su responsabilidad. Esto debilita la legitimidad pública de la decisión judicial.

Por ende, un reglamento que retenga la discrecionalidad soberana del juez y garantice el derecho a la revisión humana es algo más que un formalismo: **es la defensa de la dimensión humanista del proceso**, inseparable de la convivencia democrática. Sin tal garantía, la aplicación ciega de algoritmos corre el riesgo de derivar en resoluciones uniformes, sesgadas o descontextualizadas, en las que la persona devenga meramente objeto de categorizaciones estadísticas.

2.9.4. El “Efecto Bruselas” y el Posible Coste Competitivo

La UE ha demostrado su **capacidad de irradiar** normativas globales: sucedió con el Reglamento General de Protección de Datos (RGPD), que impulsó a multitud de empresas de Estados Unidos o Asia a reconfigurar sus políticas. Con la IA, el escenario parece repetirse. **En la administración de justicia, cualquier multinacional que desee introducir productos algorítmicos para la UE se verá obligada a cumplir con los requisitos del AI Act.** Esta presión extraterritorial, si bien asegura los derechos de los justiciables europeos, también produce:

- **Un mayor umbral de entrada:** las compañías emergentes, con recursos limitados, podrían optar por mercados sin apenas control de IA,
- **Tensiones diplomáticas:** Estados Unidos ha mostrado reticencias al ver cómo el “modelo europeo” obliga a sus gigantes tecnológicos a rendir cuentas ante la Comisión. China, por su parte, opera con su propia lógica centralista, de modo que la convergencia con Bruselas es reducida.

Aun así, la UE estima que la defensa de los derechos y libertades no es negociable, ni puede supeditarse a meros cálculos de ventaja competitiva. Antes bien, la solidez de un mercado europeo basado en la confianza y la certidumbre legal podría, a medio plazo, propiciar un crecimiento más estable de la IA.

2.9.5. Innovar sin Abdicar de la Prudencia: Cauces de Convergencia

Aun cuando la tensión entre regulación y dinamismo tecnológico persiste, **existen vías de convergencia** que pueden desactivar los temores a una paralización del progreso:

1. **Sandboxes regulatorios:** mecanismos de experimentación supervisada, donde empresas y tribunales prueban soluciones de IA en determinados casos. Estas “incubadoras jurídicas” permiten detectar fallos y sesgos antes de la adopción masiva²⁷¹.
2. **Códigos de conducta específicos** para IA judicial: el AI Act alienta la elaboración de códigos que faciliten la autodisposición empresarial a comprometerse con estándares éticos. Ello reduce burocracia y refuerza la autorregulación sin prescindir del control público (arts. 40 y 41 del AI Act).
3. **Certificaciones de conformidad:** análogas a las que surgen en materia de ciberseguridad, los proveedores podrían voluntariamente someterse a una auditoría de su IA para conseguir un sello de garantía europeo. Eso infunde confianza a los operadores jurídicos.
4. **Formación intensiva de jueces y funcionarios:** sin una alfabetización adecuada, el sistema judicial podría ser presa de la opacidad algorítmica. Es clave capacitar a los

²⁷¹ Tambiama Madiega y Anne Louise Van De Pol, *Artificial Intelligence Act and Regulatory Sandboxes: Summary*, EPoS | European Parliamentary Research Service, Members' Research Service, Informe PE 733.544, junio de 2022. Disponible en: [https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPoS_BRI\(2022\)733544_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPoS_BRI(2022)733544_EN.pdf)

magistrados en la lógica de los modelos predictivos, la identificación de sesgos y la interpretación de la evidencia generada por la IA.

Estas estrategias tienen como meta final la **coexistencia virtuosa** entre un control garantista y un clima propicio para la creatividad y la competitividad empresarial. De este modo, la administración de justicia no quedaría rezagada, sino que se beneficiaría de la IA con la certidumbre de que la esencia humana del juzgar permanece incólume.

2.9.6. La IA futura en la Justicia: Escenarios Prospectivos

A la luz de los desarrollos recientes, pueden esbozarse distintos escenarios de evolución de la IA en la justicia europea:

1. **Crecimiento progresivo de la automatización en asuntos simples:** el monitoreo y la decisión de casos fáciles—como procesos monitorios o reclamaciones de escasa cuantía—podría delegarse, en gran medida, a algoritmos controlados por un supervisor humano. El juez revisaría las resoluciones solo en caso de impugnación. Esta automatización parcial mejoraría la eficiencia sin comprometer la equidad.
2. **Auge de la “justicia asistida”:** más que sustituir al juez, los sistemas de IA servirían de apoyo para extraer patrones jurisprudenciales, cuantificar indemnizaciones o recomendar penas dentro de un rango preestablecido. El magistrado conservaría la última palabra, pero con un “radar” basado en big data y aprendizaje automático.
3. **Resistencia y heterogeneidad:** ciertos ámbitos—como el penal o el de familia—podrían mostrar una alta resistencia a la automatización, por temor a cosificar el análisis de factores subjetivos (arrepentimiento, compasión, necesidades familiares). De ahí surgiría un mosaico, con áreas de uso intensivo de IA y otras en las que prevalezca el método tradicional.
4. **Hiperrregulación o “moratoria preventiva”:** de producirse incidentes graves (e.g., sistemas de IA judicial que discriminen sistemáticamente por etnia o género), la reacción de la opinión pública podría llevar a **moratorias** o prohibiciones de amplio espectro, con efecto negativo en la investigación e implantación de la IA. Aun siendo un escenario menos deseable, no puede descartarse.

La interacción de estos escenarios dependerá de factores políticos, de la aceptación ciudadana y de la capacidad de diseñar algoritmos confiables y transparentes. Si las autoridades europeas logran que la sociedad perciba la IA judicial como un refuerzo de la igualdad ante la ley y no como un sustituto inhumano, la **adhesión y legitimidad** aumentarán.

2.9.7. El Rol de la Interpretabilidad Mecanicista y la Investigación de Vanguardia

Los avances recientes en **interpretabilidad** ofrecen un rayo de esperanza para superar la antinomia “algoritmo opaco” vs. “transparencia absoluta”. Trabajos como el de Anthropic²⁷² sobre extracción de “features” monosemánticas de grandes modelos de lenguaje sugieren que, con las técnicas adecuadas, cabe desentrañar parcialmente los engranajes internos de una red neuronal profunda. En el contexto judicial, esta “desopacificación” mitigaría riesgos de arbitrariedad y facilitaría el control jurisdiccional.

Por otra parte, la investigación en **aprendizaje causal**—capaz de detectar correlaciones espurias y aislar factores relevantes—podría reducir los sesgos discriminatorios incrustados en los datos de entrenamiento. Esta línea de vanguardia, al implementarse con rigor, encajaría como anillo al dedo con las exigencias del AI Act de proveer explicaciones razonadas y de examinar la “justicia algorítmica”²⁷³.

Si la UE impulsa y financia estas líneas de investigación, la comunidad científica europea ganará en competitividad y ofrecerá soluciones concretas a las dudas más álgidas de la IA judicial.

2.9.8. Más Allá de Europa: la Justicia Algorítmica como Debate Global

Las controversias sobre la IA en la justicia no son patrimonio exclusivo de la UE. En Estados Unidos, el uso de COMPAS en sentencias penales, o las iniciativas de “Online Dispute Resolution” con matices automatizados, han generado polémica en torno a la discriminación y la

²⁷² Adly Templeton et al., "Scaling Monosematicity: Extracting Interpretable Features from Claude 3 Sonnet," Anthropic (21 de mayo de 2024). <https://transformer-circuits.pub/2024/scaling-monosematicity/index.html>

²⁷³ Richard Oubělický, "Europe and AI: Causes and Implications of Europe Losing Ground in the Race for AI", p. 35-50.

fiabilidad. En China, los “Smart Courts” y la aplicación intensiva de la IA plantean el interrogante de si la celeridad y la uniformización priman por encima de las garantías procesales.

En este orden, la **perspectiva europea**—con su impronta garantista—cobra aún mayor relieve como contrapeso a posibles derivas tecnocráticas, ya provengan de jurisdicciones que centralizan el desarrollo de esta tecnología al son estatal (China) o de países que confían excesivamente en la corrección de la máquina por la máquina misma (Estados Unidos). El **AI Act** y la Carta Ética Europea sobre el Uso de la IA en la Justicia (CEPEJ 2018) podrían marcar un precedente normativo en foros internacionales, incidiendo en la futura “gobernanza global de la IA judicial”.

2.9.9. Conclusiones: la Senda Europea hacia una Justicia Aumentada, No Reemplazada

El despliegue de la IA en la justicia europea **no es una quimera** ni un mero experimento académico: es una realidad incipiente, con visos de expansión. Ahora bien, la UE reconoce que la automatización sin cortapisas—por “eficiente” que parezca— se arriesga a transmutar el juicio en una operación de cálculo, ajena al sustrato humano y moral que define la función judicial. De ahí el tejido de normas y cautelas que, en conjunto, aspiran a **proteger la autonomía judicial, la transparencia y la equidad**.

Las próximas décadas serán cruciales para afinar este marco. Por un lado, la UE anhela no quedar rezagada en la pugna por la innovación, por lo que incentiva laboratorios de pruebas, certificaciones y la investigación en interpretabilidad. Por otro lado, sostiene sin titubeos la línea roja de la dignidad y los derechos fundamentales: **ningún auge tecnológico, por colosal que sea, puede legitimar la indefensión del ciudadano ni la abdicación de valores constitucionales**.

La “tercera vía garantista” es, en definitiva, una apuesta por la **justicia aumentada**: algoritmos que allanen procesos y perfeccionen la gestión, pero sin destronar la razón prudencial ni trivializar la singularidad de cada caso. Lejos de constituir un “freno” estéril, la regulación europea se postula como **el anclaje** que salva a la justicia de la deshumanización y la transforma, con la ayuda de la IA, en una institución más ágil y accesible, pero siempre fiel al espíritu de la democracia constitucional. Ese es el gran desafío que, con rigor y prudencia, la UE se propone encarar en los años venideros.

Capítulo III: La Implementación de Sistemas de IA en la Administración de Justicia Costarricense: Diagnóstico, Desafíos y Propuestas desde la Experiencia Europea

3.1.- Estado Actual del Marco Normativo Costarricense

3.1.1.- Análisis del Marco Constitucional

La irrupción de los sistemas de Inteligencia Artificial (IA) en el ámbito de la administración de justicia suscita un profundo debate sobre su encaje constitucional, habida cuenta de las implicaciones que estas tecnologías disruptivas pueden tener sobre principios basilares del Estado de Derecho como la independencia judicial, la tutela judicial efectiva o la igualdad ante la ley. Un análisis exhaustivo de esta cuestión exige, pues, escrutar minuciosamente el texto de nuestra Carta Magna para identificar aquellos preceptos que, ya sea de forma expresa o implícita, proyectan su imperio sobre el uso de algoritmos en la función jurisdiccional.

➤ Independencia judicial (art. 9 y 154)

El principio de independencia judicial, cincelado en el frontispicio del Estado constitucional de Derecho como condición de posibilidad de su misma existencia, atraviesa en el momento presente una suerte de "prueba ácida" ante los desafíos suscitados por la irrupción de la inteligencia artificial en el sanctasanctórum de la administración de justicia²⁷⁴.

La profundidad de las transformaciones que estas tecnologías disruptivas anuncian en la morfología y fisiología de la función jurisdiccional exige de la ciencia jurídica un renovado esfuerzo de revisitación de sus fundamentos y resignificación de su alcance ante lo que se ha dado en llamar la "cuarta revolución industrial"²⁷⁵. Un repensar de calado el principio basilar que, lejos

²⁷⁴ Silvia Barona Vilar, "Dataización de la justicia (algoritmos, inteligencia artificial y justicia, ¿el comienzo de una gran amistad?) / Datization of Justice (Algorithms, Artificial Intelligence und Justice, The Beginning of a Great Friendship?)," Revista Boliviana de Derecho, no. 36 (julio 2023): 14–45. ISSN: 2070-8157. Recuperado de: <https://dialnet.unirioja.es/descarga/articulo/9043836.pdf>

²⁷⁵ Klaus Schwab, *La cuarta revolución industrial* (The Fourth Industrial Revolution) Ginebra: Foro Económico Mundial, 2016; edición en español, México: Penguin Random House Grupo Editorial, S. A. de C. V., 2017, p. 12. Recuperado de: <https://economiapoliticaenam.wordpress.com/wp-content/uploads/2020/05/klaus-schwab.la-4c2b0-rev.-industrial-2.pdf>

de fosilizarse en la nostalgia de planteamientos pretéritos, sepa proyectar su vis expansiva garantista a las inéditas tesituras de una "justicia aumentada".

Ciertamente, el advenimiento de sistemas expertos capaces de procesar ingentes volúmenes de datos jurídicos, "aprender" de la experiencia acumulada en los repertorios jurisprudenciales y orientar o incluso sugerir cursos decisorios al juzgador humano, plantea una panoplia de interrogantes de primer orden para la dogmática ius publicista. Desde el riesgo de atrofia o jibarización de la "conciencia jurídica" del juez, reducido a mero usuario o "validador" del criterio apodíctico de la máquina²⁷⁶ hasta el pavor reverencial a los "oráculos algorítmicos", pasando por el fantasma de la homogeneización o "isonomía computacional" de las decisiones, los puntos de fricción entre la IA judicial y el reducto irreductible de autonomía valorativa que inerva el principio constitucional de independencia son múltiples y variopintos.

Ahora bien, si huyendo de apocaliptismos simplistas reconducimos esta cuestión crucial a sus justos términos de análisis jurídico ponderado, habremos de convenir en que el constitucionalismo contemporáneo provee un arsenal de herramientas metodológicas y axiológicas válidas para un encaje solvente de las utilidades de la "justicia digital" en el marco infranqueable de los principios y derechos que cimientan el ideal del "gobierno de las leyes". Lejos de propugnar una suerte de "neo-ludismo judicial" contrario a toda virtualidad de una administración de justicia tecnológicamente avezada, de lo que se trata es de pergeñar soluciones innovadoras que permitan conciliar las ventajas incuestionables de la automatización y la analítica predictiva con la preservación del núcleo indisponible de discrecionalidad judicial en que se cifra toda esperanza de Justicia situada.

Para abordar este desafío titánico con el rigor técnico-jurídico que reclama, no podemos obviar la densidad conceptual y la pluridimensionalidad de un principio tan proteico y multiforme como el de la independencia judicial. Tal y como ha recordado en fechas recientes la Comisión Venecia, la independencia no se configura exclusivamente como una prerrogativa o atributo del estatuto personal del juez, sino que puede y debe ser también considerada como un derecho fundamental de los ciudadanos y como una garantía institucional de un poder del Estado²⁷⁷. Esta

²⁷⁶ Andrea Simoncini, "L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà," BioLaw Journal – Rivista di BioDiritto 1, no. 1 (2018): p. 81. Recuperado de: https://www.academia.edu/94336762/L_algoritmo_incostituzionale_intelligenza_artificiale_e_il_futuro_delle_libert%C3%A0

²⁷⁷ Comisión Europea para la Democracia a través del Derecho (Comisión de Venecia), Informe preliminar sobre la independencia del sistema judicial: Parte I: La independencia de los jueces, Estudio No. 494/2008, Estrasburgo, 5 de

doble naturaleza jurídica, subjetiva por su directa vinculación con el derecho a la tutela judicial efectiva, pero al tiempo objetiva por su imbricación con la cláusula estructural de separación de poderes, permea toda reflexión constitucional sobre el alcance del principio y dota a las garantías que lo instrumentan de una singular vis atractiva hermenéutica.

Como punto de partida inexcusable, debemos recordar que la independencia judicial encuentra su formulación más prístina en el artículo 154 de la Constitución Política, a cuyo tenor "el Poder Judicial sólo está sometido a la Constitución y a la ley". Esta lacónica proclamación, condensa en su aparente simplicidad todo un haz de implicaciones normativas y de profundas resonancias iusfilosóficas. Y es que, al postular la vinculación exclusiva de los tribunales al bloque de la legalidad constitucional, el precepto excluye por principio cualquier subordinación de la función jurisdiccional a coordenadas extrajurídicas, ya sean de índole política, social o económica. La aplicación del Derecho se erige así en la única brújula axiológica del juez, excluyendo injerencias espurias en el sagrado recinto de la potestad de juzgar y hacer ejecutar lo juzgado. Dicha proclama no se agota en dicha norma.

Antes bien, este principio medular aparece revestido de la condición de estándar iusfundamental a la vez por su consagración en los principales tratados internacionales sobre derechos humanos ratificados por Costa Rica, destacadamente la Declaración Universal (art. 10), el Pacto Internacional de Derechos Civiles y Políticos (art. 14) o la Convención Americana (art. 8). Esta "convencionalización" del principio compele a una lectura de sus exigencias conforme al acervo interpretativo del Derecho Internacional de los Derechos Humanos del que son piezas angulares los Principios Básicos sobre Independencia de la Judicatura²⁷⁸, la jurisprudencia de la Corte Interamericana²⁷⁹ o los criterios generales de aplicación de la cláusula del "juez natural" destilados por el Comité de Derechos Humanos²⁸⁰.

marzo de 2010, CDL(2010)006, p. 3. Recuperado de:
[https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL\(2010\)006-e](https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL(2010)006-e)

²⁷⁸ Naciones Unidas, Principios básicos relativos a la independencia de la judicatura, 1985. Recuperado de: <https://www.ohchr.org/es/instruments-mechanisms/instruments/basic-principles-independence-judiciary>

²⁷⁹ Corte Interamericana de Derechos Humanos, *Caso Apitz Barbera y otros ("Corte Primera de lo Contencioso Administrativo") vs. Venezuela*, Sentencia de 5 de agosto de 2008, parr. 55. Recuperado de: https://www.corteidh.or.cr/docs/casos/articulos/seriec_182_esp.pdf

²⁸⁰ Comité de Derechos Humanos, Observación general núm. 32, artículo 14: El derecho a un juicio imparcial y a la igualdad ante los tribunales y cortes de justicia, 2007, párr. 19. Recuperado de: <https://www.refworld.org/es/leg/coment/ccpr/2007/es/52583>

A las aportaciones de este "bloque de constitucionalidad" *lato sensu*, debemos añadir los sustanciales desarrollos jurisprudenciales del principio acometidos tanto por nuestros tribunales domésticos²⁸¹ como por instancias regionales de fiscalización iusfundamental como la Corte Europea de Derechos Humanos²⁸². De la interacción dialógica entre los estándares decantados en Estrasburgo y las tradiciones constitucionales de los Estados miembros del Consejo de Europa ha emergido un denso y matizado corpus de criterios y subprincipios que dan concreción y efectividad al mandato genérico de independencia judicial: desde las facetas de inamovilidad, preconstitución legal o autonomía decisoria como atributos medulares del derecho al "juez imparcial"²⁸³ hasta la doctrina del "tribunal independiente" como piedra siller de las garantías del efectivo acceso a la jurisdicción y el debido proceso²⁸⁴.

De este ingente acervo normativo, jurisprudencial y de *soft law* (ver apartado sobre la Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales y su Entorno) que integra el canon hermenéutico desde el cual debemos examinar los nuevos desafíos tecnológicos, conviene extraer tres vectores maestros a cuya luz calibraremos los límites y las posibilidades del uso de sistemas inteligentes en los procesos judiciales.

El primero de ellos ataña a la propia configuración iusfilosófica del principio de independencia y su evolución desde una concepción puramente formal o institucional hacia una visión material o axiológica. En efecto, la emancipación de la judicatura frente a las injerencias de los otros poderes, especialmente del ejecutivo, se concibió inicialmente como una garantía enderezada a asegurar la mera "subsunción mecánica del Derecho" por un aplicador fungible y despersonalizado. La imagen montesquiana del juez como "*bouche de la loi*" refleja prístinamente ese ideal ilustrado de una magistratura políticamente neutralizada mediante la servidumbre a un

²⁸¹ Sala Constitucional de la Corte Suprema de Justicia de Costa Rica, voto N.º 4849-2009 de las 13:17 horas del 20 de marzo de 2009, y voto N.º 5790-99 de las 16:21 horas del 11 de agosto de 1999.

²⁸² Tribunal Europeo de Derechos Humanos, Caso Maktouf y Damjanović contra Bosnia y Herzegovina, sentencia de 18 de julio de 2013, párr. 49. Recuperado de: [https://hudoc.echr.coe.int/fre#%22itemid%22:\[%22002-4870%22\]}](https://hudoc.echr.coe.int/fre#%22itemid%22:[%22002-4870%22]})

²⁸³Ibid.

²⁸⁴ Tribunal Europeo de Derechos Humanos, *Caso Baka contra Hungría*, sentencia de 23 de junio de 2016, párr. 73. Recuperado de: [https://hudoc.echr.coe.int/eng#%22itemid%22:\[%22001-144139%22\]}](https://hudoc.echr.coe.int/eng#%22itemid%22:[%22001-144139%22]})

ordenamiento que se postula como semánticamente autosuficiente y que no deja resquicios a la valoración personal.

Esta concepción formalista-mecanicista, consustancial al positivismo decimonónico, pronto evidenció sus fisuras ante la inexorable "apertura" de los sistemas jurídicos a valores y principios supralegales. La "revuelta contra el formalismo" que se opera en la teoría jurídica de la primera mitad del siglo XX, merced al impulso de corrientes como la jurisprudencia de intereses, el realismo jurídico o el neouisnaturalismo y que alcanza su epítome con la consolidación del paradigma neoconstitucionalista en la segunda posguerra²⁸⁵ abonará una visión mucho más densa y articulada del rol judicial. La plenitud hermética del derecho legal dará paso a la textura abierta de normas principales y cláusulas generales cuya aplicación demanda necesariamente una labor creativa de "desarrollo judicial del Derecho". Los principios de proporcionalidad, razonabilidad y efecto útil emergen como puentes hermenéuticos para conjugar los fines del sistema jurídico con las circunstancias del caso concreto.

Esta evolución, que supone una potenciación del activismo judicial no implica, sin embargo, una "capitulación decisionista" en manos de un juez "cadí" o soberano. Antes bien, la judicialización de la producción normativa se concibe siempre como un razonamiento disciplinado por pautas metódicas de argumentación jurídica y por procedimientos discursivos de legitimación democrática de las sentencias. Los desarrollos de la segunda mitad del siglo XX en el campo de la hermenéutica filosófica, la tópica jurídica o las teorías de la argumentación proveerán el herramiental metodológico necesario para reconducir el momento valorativo o "creativo" de la decisión judicial a cánones rigurosos y controlables de racionalidad práctica. Las exigencias de consistencia, coherencia, universalidad, consecuencialismo y respeto de precedentes actúan como límites inmanentes a la discrecionalidad del intérprete, reconduiéndola a estándares de motivación reforzada que exigen justificar en términos de "razón pública"²⁸⁶ las decisiones adoptadas, especialmente cuando estas se apartan de la literalidad de la norma o entran en fricción con bienes o derechos fundamentales.

²⁸⁵ Gustavo Zagrebelsky, *El derecho dúctil: Ley, derechos, justicia*, traducción de Marina Gascón, Madrid: Editorial Trotta, 2011, p. 93-97. Recuperado de: https://www.academia.edu/4980303/155026921_El_Derecho_Ductil_Gustavo_Zagrebelsky.pdf

²⁸⁶ John Rawls, *La justicia como equidad: Una reformulación*, edición a cargo de Erin Kelly (Barcelona: Paidós, 2001), 247-254. Recuperado de: <https://www.terrileyasociados.com.ar/post/john-rawls-la-justicia-como-equidad-una-reformulacion-a-cargo-de-erin-kelly-paidos.pdf>

Es en este contexto epistemológico donde debemos incardinarn la radical transformación que ha experimentado el alcance del principio de independencia judicial por mor de la constitucionalización del ordenamiento jurídico. Al erigirse la Carta Magna en fuente suprema de validez no solo formal sino también material del sistema normativo, proveyendo una "ductilidad" o maleabilidad axiológica que permite maximizar dinámicamente los principios fundamentales, la sumisión del juez al "imperio de la ley" se transmuta en vinculación prioritaria a los valores y derechos constitucionalmente consagrados. Se instaura así una "legalidad constitucional" que trasciende la estrecha subsunción normativa para abarcar un juicio de "razonabilidad práctica" en el que se sopesan circunstancias singulares, consecuencias aplicativas y exigencias tópicas de equidad sustancial.

Como corolario de esta evolución, el principio de independencia adquiere una dimensión sustantiva o valorativa, concibiéndose no ya como mera ausencia de vínculos imperativos con los otros poderes sino como garantía positiva de sujeción del juez a un orden constitucional de valores que ha de salvaguardar frente a cualquier intento de instrumentalización política. La "libertad interpretativa" inherente al ejercicio de la jurisdicción, lejos de traducirse en decisionismo o arbitrariedad, se encauza hacia la "única respuesta correcta" que pondere equilibradamente los principios fundamentales en juego, primando la "justicia del caso concreto" sobre automatismos subsuntivos. La independencia deviene, en suma, presupuesto y al tiempo consecuencia de la argumentación constitucional, pues solo desde la autonomía axiológica puede el juez armonizar la universalidad de las normas con la irrepetibilidad de cada litigio, pero al hacerlo queda vinculado a un deber reforzado de motivación racional en el que cifra su propia legitimidad democrática.

Es este entendimiento "post-positivista" de la independencia judicial como reserva de jurisdicción el que debe proyectarse al despliegue de sistemas inteligentes en los procesos judiciales, ponderando reflexivamente sus ventajas e identificando cautelas frente a eventuales erosiones del principio. Desde esta óptica sustantivista, la "gobernanza algorítmica" de la función jurisdiccional solo resultará admisible en la medida en que potencie la calidad argumentativa de las decisiones sin cercenar la capacidad del juez de "decir el derecho" desde valoraciones circunstanciadas. Los modelos que releguen al juzgador a mero aplicador mecánico de las correlaciones arrojadas por el software, obviando el "momento" prudencial de individuación equitativa, serán difícilmente compatibles con un diseño constitucional que hace de la independencia un reducto inexpugnable de garantía ciudadana y no una mera proclama huera.

Junto con este escrutinio de legitimidad constitucional, el despliegue judicial de la IA reclama un marco regulatorio garantista que, desde la reserva de ley, establezca presupuestos habilitantes, cautelas procedimentales y reglas de imputación de responsabilidad claramente definidos. El principio de independencia judicial exige que sea el legislador democrático quien acote los confines de lo jurídicamente admisible, estableciendo una nítida diferenciación entre aquellos usos puramente "logísticos" o de asistencia documental, cuya implementación puede confiarse al nivel reglamentario, y aquellos otros que afecten a facetas nucleares de la función jurisdiccional y que, por tanto, demandan una disciplina legal tanto en la predeterminación de su ámbito de aplicación como en la articulación de salvaguardas frente a injerencias indebidas.

En esta línea, no parece aventurado postular una suerte de principio de "intervención o supervisión humana" como presupuesto habilitante para la validez constitucional del uso de IA en entornos judiciales, el cual se materializaría en exigencias graduadas de control por el juez de las decisiones automatizadas que pudieran incidir en la esfera de derechos de los justiciables. Desde la transparencia algorítmica que haga cognitivamente accesible el "cómo" y el "porqué" de las inferencias empleadas, hasta el reconocimiento de un auténtico "derecho de impugnación" que permita revisar y, en su caso, corregir los eventuales sesgos o disfunciones de la máquina, pasando por estándares reforzados de motivación que expliciten el peso conferido a las orientaciones del software en el proceso deliberativo, este haz de cautelas respondería a la lógica de situar al ser humano en el centro de los procesos decisorios algorítmicamente mediados.

De particular relevancia en este contexto resulta delimitar con tino el nexo de imputación de responsabilidad por los daños eventualmente irrogados por resoluciones írritas o desproporcionadas adoptadas con apoyo de IA. Si bien el instituto de la responsabilidad patrimonial del Estado por error judicial o funcionamiento anormal de la administración de justicia brinda un cauce reparador de las lesiones antijurídicas sufridas por los justiciables²⁸⁷, su

²⁸⁷ En ese sentido, ver el **voto No. 2024-4230** emitido por el Tribunal Contencioso Administrativo y Civil de Hacienda a las ocho horas con tres minutos del uno de julio de 2024. La responsabilidad por función judicial en Costa Rica encuentra su fundamento constitucional en los artículos 9, 11, 33, 41 y 154 de la Carta Magna, configurándose como un régimen de responsabilidad objetiva del Estado-Juez que surge tanto por error judicial como por funcionamiento anormal de la administración de justicia. Este sistema se consolidó jurisprudencialmente a partir del voto 5981-95 de la Sala Constitucional, que estableció que la responsabilidad del Estado derivada del ejercicio de la función jurisdiccional no requiere desarrollo legislativo específico para su procedencia, pues emana directamente de los principios constitucionales que garantizan la responsabilidad del Estado y la tutela judicial efectiva. En cuanto a su configuración específica, el régimen distingue entre la responsabilidad derivada de la función jurisdiccional propiamente dicha y aquella que surge de la función administrativa del Poder Judicial, aplicándose a esta última las normas sobre responsabilidad de la Ley General de la Administración Pública. En materia penal, el artículo 271 del

efectividad puede verse comprometida por la opacidad o inescrutabilidad de los sistemas expertos, así como por la concurrencia de actores públicos y privados en su diseño e implementación. De ahí la importancia de establecer ex ante protocolos claros de evaluación de riesgos, auditabilidad e interoperabilidad de los modelos adquiridos por la administración de justicia, acompañados de exigentes cláusulas de compliance tecnológico a los proveedores que disuadan estrategias de irresponsabilidad organizada.

Desde el punto de vista disciplinario, la independencia judicial podría verse paradójicamente reforzada por la fijación de estándares normativos de diligencia en el uso y la supervisión de herramientas inteligentes por parte de jueces y magistrados. El deber de adoptar decisiones informadas y motivadas, núcleo irreductible de la independencia como responsabilidad, cobraría así una dimensión tecnológica, compeliendo al juzgador a emplear razonable y críticamente las asistencias algorítmicas a su alcance, so pena de incurrir en infracciones deontológicas

Finalmente, un tercer vector de inexcusable ponderación constitucional nos sitúa ante la aspiración de estándares globales o armonizados para el gobierno de la IA judicial en clave cosmopolita y de Derecho comparado. En este contexto, iniciativas pioneras como el ya referido Ethical Charter on the Use of AI in Judicial Systems auspiciado por la Comisión Europea para la Eficacia de la Justicia, permiten atisbar los mimbres de una suerte de "orden público tecnológico" que, desde el prisma tuitivo de los derechos fundamentales, proporcione un marco de controlabilidad del acelerado proceso de plataformización de la justicia a escala global.

En esa misma dirección camina, con un calado jurídico-positivo muy superior el Reglamento sobre Inteligencia Artificial de la Unión Europea cuyo articulado contempla previsiones específicas para preservar la independencia judicial en entornos automatizados. Así, además de incluir expresamente los usos forenses de IA en la categoría de "alto riesgo", sometidos a requisitos reforzados de transparencia, explicabilidad, trazabilidad y supervisión humana, el texto prohíbe el empleo de sistemas automatizados para adoptar resoluciones finales sobre el fondo del litigio, circunscribiendo su lícito aprovechamiento a las funciones de apoyo a la valoración judicial de pruebas y hechos.

Código Procesal Penal establece condiciones particulares para la procedencia de la responsabilidad estatal, requiriendo la demostración de arbitrariedad o culpa grave del funcionario para medidas cautelares en general, y exigiendo la plena demostración de inocencia -no bastando la simple absolutoria por duda razonable- en casos de prisión preventiva, configurándose así un régimen especial que delimita la responsabilidad del Estado-Juez en el ámbito penal.

Esta apuesta decidida por salvaguardar el "momento humano" de la decisión jurisdiccional, relegando a los algoritmos inteligentes a un rol meramente "servicial" o de asistencia, refleja una concepción garantista de la IA judicial plenamente coherente con los principios axiológicos del constitucionalismo europeo, y muy señaladamente con la centralidad de la "dignidad humana" como premisa antropológica fundamental. Dignidad que, al cifrar la condición de persona en su capacidad de autodeterminación moral, veda por principio cualquier cesión incondicionada de la responsabilidad decisoria a sistemas expertos, por muy sofisticados y fiables que sean sus modelos predictivos o sus protocolos de autoaprendizaje.

Ciertamente, los ingentes retos regulatorios, interpretativos y aplicativos que afrontará el despliegue del Reglamento Europeo de IA en el ámbito judicial —desde la concreción de las "evaluaciones de conformidad" para acreditar el cumplimiento de las exigencias de gobernanza de datos, registro de eventos o ciberseguridad, hasta la articulación de mecanismos efectivos de supervisión y ejecución a cargo de "autoridades nacionales de control"— están aún lejos de quedar despejados. Con todo, la inequívoca vocación armonizadora del instrumento y su reforzada condición de aplicabilidad directa como reglamento comunitario permiten colegir un impacto muy significativo en los ordenamientos domésticos, forzando adaptaciones normativas y cambios culturales en la judicatura de los Estados miembros.

No será tarea fácil conjugar, en el contexto de administraciones de justicia aún lastradas por la precariedad de medios y la obsolescencia tecnológica, las comprensibles demandas de agilidad y eficiencia que alientan el recurso a soluciones automatizadas con los inderogables requerimientos de independencia e imparcialidad que la tutela judicial efectiva demanda. Quizás el gran desafío, no sea solo cómo "regular la tecnología", sino sobre todo cómo "tecnificar el Derecho" sin disolver su médula garantista. Una regulación constitucionalmente adecuada de la IA judicial habrá de partir, pues, de un enfoque dinámico y relacional de la independencia, que no la conciba como una prerrogativa adquirida o un privilegio de casta, sino como un principio vivificador en constante interacción con las cambiantes realidades sociotécnicas.

Concluyendo, el impacto de la IA en el principio constitucional de independencia judicial reviste una trascendencia indisputada, habida cuenta de su imbricación profunda con la cláusula básica del Estado de Derecho y su virtualidad práctica como garantía de los derechos fundamentales en el proceso. La creciente sofisticación de los modelos computacionales predictivos o suasivos y su proyección a ámbitos sensibles de la función jurisdiccional, desde la

admisión y tramitación de las demandas hasta la motivación y revisión de las sentencias, pasando por la valoración de pruebas o la evaluación de riesgos cautelares, suscita complejos desafíos regulatorios, interpretativos y aplicativos para salvaguardar el núcleo irreductible de autonomía decisoria que es condición necesaria de legitimación democrática de la judicatura.

Sin desconocer los esperanzadores frutos de eficiencia, coherencia y calidad que cabe esperar de un uso constitucionalmente orientado de los sistemas inteligentes, de lo que se trata es de modular su inserción en los procesos judiciales desde una nítida delimitación entre aquellos aspectos fungibles o rutinarios que admiten una automatización controlada y aquellos otros que afectan a elementos basilares de la potestad de juzgar y hacer ejecutar lo juzgado, respecto de los cuales la supervisión humana debe reputarse un presupuesto habilitante inexcusable. El recurso a modelos explicativos que hagan accesible la lógica algorítmica subyacente a las decisiones, la articulación de vías impugnatorias que aseguren un debido escrutinio racional de los sesgos eventualmente discriminatorios de la IA y el anclaje riguroso de su despliegue a reservas de ley parlamentaria que preserven la preeminencia del Derecho y los derechos, emergen como exigencias medulares de un marco regulatorio garantista a escala nacional y transnacional.

Solo desde esta óptica de "optimización ponderada" de la eficiencia tecnológica y la legitimidad institucional, que conciba al juez como "consumidor proactivo" y "auditor cualificado"—pero en ningún caso "esclavo complaciente"— de las herramientas inteligentes a su alcance, será posible transitar el profundo cambio de paradigma que la IA augura en la administración de justicia preservando la quintaesencia constitucional de la independencia judicial. Una independencia que, lejos de concebirse como repliegue defensivo o aversión tecno-escéptica, debe reformularse en clave relacional como "interdependencia" de la función jurisdiccional con su contexto científico-social, como apertura reflexiva a la innovación que no abdique de su compromiso esencial con la garantía de los derechos de la persona.

El reto es mayúsculo, qué duda cabe, pero las coordenadas axiológicas de nuestra tradición constitucional y muy particularmente el principio de dignidad humana como premisa antropológica, proporcionan una brújula segura para surcar las procelosas aguas de la disruptión digital sin zozobrar en la anegación de los más altos valores del ordenamiento jurídico

➤ **Debido Proceso**

Un segundo bloque de normas constitucionales que reviste especial trascendencia en el análisis del impacto de la inteligencia artificial en la administración de justicia es el relativo a las garantías jurisdiccionales y procesales de los justiciables. Nuestra Carta Magna, con el loable propósito de proscribir cualquier atisbo de arbitrariedad en el ejercicio de la potestad judicial, consagra un haz de derechos fundamentales que conforman el estatuto básico de toda persona que se ve compelida a comparecer ante los tribunales, ya sea en calidad de demandante o demandado, acusador o acusado.

El epicentro de este entramado garantista, auténtica clave de bóveda del Estado de Derecho en su dimensión procesal, lo encontramos en el artículo 41 de la Constitución, a cuyo tenor "ocurriendo a las leyes, todos han de encontrar reparación para las injurias o daños que hayan recibido en su persona, propiedad o intereses morales. Debe hacérseles justicia pronta, cumplida, sin denegación y en estricta conformidad con las leyes". La riqueza semántica de este precepto, reverberada en cada uno de sus sintagmas, es de tal magnitud que bien podría afirmarse que compendia en su seno la quintaesencia de la tutela judicial efecto. Asimismo, el artículo 39, por ejemplo, dispone que nadie puede ser penado sino "*en virtud de sentencia firme dictada por autoridad competente, previa oportunidad concedida al indiciado para ejercitar su defensa y mediante la necesaria demostración de culpabilidad*".

Asimismo, la garantía de defensa y audiencia bilateral (derecho a ser oído) forma parte esencial del debido proceso. En la administración de justicia, **cualquier uso de IA debe respetar estas garantías procesales**. Esto significa que las partes en un proceso tienen derecho a conocer, intervenir y controvertir las decisiones o recomendaciones generadas por un sistema automatizado. Si un tribunal utiliza, por ejemplo, un algoritmo para sugerir una sentencia o calificar un riesgo procesal, las partes **deben tener oportunidad de refutar o examinar** esa información. Además, el debido proceso incluye la obligación de motivar las sentencias ("sentencia justa y exigencia de motivación"), lo cual implica que **la decisión final debe estar debidamente fundamentada en términos comprensibles**.

Un "resultado" opaco producido por una IA no exime al juez de explicar las razones del fallo. Por ende, la falta de transparencia de ciertos algoritmos (las llamadas "*cajas negras*") puede entrar en tensión con el debido proceso, ya que dificulta el control de la correcta fundamentación de la decisión. La jurisprudencia constitucional comparada ilustra este problema: en el caso *State*

v. Loomis (Wisconsin, EE. UU.)²⁸⁸, se cuestionó que un juez penal basara la sentencia en un puntaje de riesgo proporcionado por un algoritmo no transparente, alegando violación al debido proceso porque la defensa no podía examinar cómo operaba dicho algoritmo.

En resumen, el **uso de IA debe estar diseñado de forma que respete el derecho a un proceso justo**, garantizando siempre la intervención humana para salvaguardar la defensa y la motivación adecuada de las resoluciones.

➤ Igualdad ante la Ley y No Discriminación

El principio de igualdad está consagrado en el artículo 33 constitucional: “*Toda persona es igual ante la ley y no podrá practicarse discriminación alguna contraria a la dignidad humana*” Este mandato implica que las decisiones judiciales no pueden establecer tratos diferenciados injustificados. La incorporación de IA en la justicia plantea desafíos respecto de posibles **sesgos algorítmicos**. Si los datos o algoritmos utilizados contienen prejuicios (por ejemplo, por origen étnico, género, condición socioeconómica), existe el riesgo de que reproduzcan o agraven la discriminación, violando el principio de igualdad²⁸⁹ La Sala Constitucional ha desarrollado una extensa jurisprudencia protegiendo la igualdad y prohibiendo discriminaciones arbitrarias²⁹⁰, exigiendo un test estricto de razonabilidad para cualquier trato diferenciado. Por tanto, **un sistema de IA utilizado en tribunales debe ser auditado y entrenado cuidadosamente para evitar sesgos**, garantizando resultados imparciales. Experiencias comparadas muestran la relevancia de este punto: por ejemplo, en EE. UU. se descubrió que el algoritmo *COMPAS* (usado para evaluar riesgo de reincidencia) tenía a puntar más alto a acusados afrodescendientes, evidenciando un sesgo racial. Algo similar no sería admisible en Costa Rica dado el mandato de igualdad. De hecho, informes internacionales estiman que alrededor del 85 % de los proyectos de IA enfrentan

²⁸⁸ "State v. Loomis. La Corte Suprema de Wisconsin exige advertencia previa al uso de evaluaciones algorítmicas de riesgo en la determinación de sentencias (comentario sobre 881 N.W.2d 749 [Wis. 2016])," Harvard Law Review 130, no. 5 (marzo de 2017): disponible en <https://harvardlawreview.org/print/vol-130/state-v-loomis/>

²⁸⁹ UNESCO, "Herramientas para un uso ético: Jueces de América Latina y el Caribe se capacitan en Inteligencia Artificial y Estado de Derecho", Noticia, 15 de noviembre de 2023, Recuperado de: <https://es.unesco.org/news/herramientas-para-un-uso-etico-jueces-de-america-latina-y-caribe-capacitan-en-inteligencia-artificial>

²⁹⁰ Ver las siguientes sentencias de la Sala Constitucional de la Corte Suprema de Justicia en materia de igualdad: votos n.º 1770-94 y 1045-94

obstáculos debido a sesgos algorítmicos y falta de explicabilidad²⁹¹. Cualquier herramienta automatizada en el ámbito judicial costarricense debe, entonces, **prevenir la discriminación y asegurar un trato igualitario**. Normativas propuestas sobre IA en Costa Rica ya reconocen este principio, obligando a alinear el uso de IA con los valores de dignidad e igualdad humana²⁹². En suma; la Sala IV, garante de este principio, muy probablemente censuraría una aplicación tecnológica que genere decisiones parcializadas o desiguales sin justificación objetiva y razonable.

➤ Derecho a la Privacidad y Protección de Datos Personales

El ordenamiento costarricense tutela fuertemente el derecho a la intimidad y la protección de datos. El artículo 24 constitucional garantiza “*el derecho a la intimidad, a la libertad y al secreto de las comunicaciones*”, declarando inviolables los documentos privados y comunicaciones, salvo excepciones legales muy calificadas. Estas excepciones – como interceptaciones o inspecciones de datos – solo pueden autorizarse por ley aprobada por mayoría calificada y bajo control judicial estricto, siendo **indelegable la responsabilidad de la autoridad judicial** en su aplicación.

En materia de datos personales, Costa Rica cuenta además con la Ley 8968 “Protección de la Persona frente al Tratamiento de sus Datos Personales”, que desarrolla el derecho a la autodeterminación informativa. La propia Sala Constitucional ha reconocido el derecho a la privacidad y al control de los datos personales como parte de la dignidad humana, amparando por ejemplo a individuos frente a exposiciones indebidas de su información en entornos digitales. Un caso relevante es la resolución N° 2018-16787 de la Sala IV²⁹³, en la que se ordenó al Registro Civil suprimir de su sitio web el dato del sexo registral de una persona, por considerar que la divulgación en línea de esa información vulneraba su derecho a la intimidad. Este precedente evidencia cómo la Sala aplica los principios de privacidad en contexto tecnológico: la información personal en plataformas digitales oficiales debe protegerse si su publicidad innecesaria afecta derechos fundamentales. Al implementar IA en la administración de justicia, surgen varios retos

²⁹¹ Regina García, “¿La inteligencia artificial tiene sesgos?,” *Instituto Mexicano para la Competitividad A.C.*, 8 de febrero de 2023, <https://imco.org.mx/la-inteligencia-artificial-tiene-sesgos/>

²⁹² Asamblea Legislativa de la República de Costa Rica, *Proyecto de Ley N.º 23.771, Ley de Regulación de la Inteligencia Artificial en Costa Rica* (2023).

²⁹³ Sala Constitucional, *Resolución N° 16787-2018*, Expediente 18-009250-0007-CO (San José, Costa Rica, 5 de octubre de 2018).

ligados a la privacidad: las herramientas de IA suelen requerir grandes volúmenes de datos (muchos de ellos sensibles) y pueden implicar análisis masivos de expedientes judiciales o información personal de litigantes. Es imprescindible entonces garantizar el **cumplimiento estricto de las normas de protección de datos**. Cualquier sistema de IA judicial debe operar con bases de datos seguras, con datos anonimizados cuando corresponda y respetando los consentimientos y las finalidades permitidas. Un ejemplo positivo en Costa Rica es la reciente **herramienta de IA para la despersonalización de sentencias**. El Poder Judicial implementó en 2024 un sistema que anonimiza las resoluciones judiciales antes de su publicación, removiendo nombres y datos identificatorios, precisamente para dar cumplimiento a la Ley 8968²⁹⁴. Esta herramienta –desarrollada por la Comisión de Protección de Datos del Poder Judicial –**demuestra cómo la IA puede usarse para fortalecer la privacidad**, asegurando la tutela de los derechos de las personas usuarias de la justicia al publicar jurisprudencia sin exponer datos personales.

En resumen, el marco constitucional exige que la IA en el ámbito judicial **no invada la esfera privada** de las personas: debe haber confidencialidad de la información sensible y respeto por el *habeas data*. La transparencia de la gestión judicial no puede confundirse con violación de la intimidad: la Sala IV probablemente invalidaría cualquier uso de IA que genere tratamientos masivos de datos personales sin base legal o que comprometa la privacidad de los involucrados en un proceso.

➤ Seguridad Jurídica

La seguridad jurídica es un principio constitucional implícito que la Sala Constitucional ha desarrollado como parte del Estado de Derecho²⁹⁵. Implica certidumbre, predictibilidad y estabilidad en la aplicación de la ley, de modo que las personas puedan confiar en que sus derechos serán respetados y las reglas del juego no cambiarán arbitrariamente.

²⁹⁴ Poder Judicial de Costa Rica, “Novedosa herramienta de Inteligencia Artificial se aplica en mejora de la protección de datos,” presentado el 20 de marzo de 2024. Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/novedosa-herramienta-de-inteligencia-artificial-se-aplica-en-mejora-de-la-proteccion-de-datos?catid=8&Itemid=409>

²⁹⁵ Voto No. 8390-97 emitido por la Sala Constitucional de la Corte Suprema de Justicia a las 16:21 hrs del 9 de diciembre de 1997.

En la administración de justicia, la seguridad jurídica se manifiesta en la exigencia de decisiones coherentes, aplicación uniforme de la ley, respeto por la cosa juzgada y, en general, en un funcionamiento predecible y ordenado del aparato judicial. La introducción de sistemas de IA podría tensar este principio si no se hace cuidadosamente. Por ejemplo, un algoritmo de decisión judicial no transparente podría minar la certidumbre en los procesos: las partes podrían desconocer los criterios con que se decide su caso, u obtener resultados inconsistentes. Para proteger la seguridad jurídica, **la IA empleada debe ser comprensible, confiable y producir resultados reproducibles bajo las mismas condiciones**. La Sala Constitucional ha señalado que la exigencia de motivación de las sentencias es una condición de la seguridad jurídica y de la justicia efectiva – los justiciables deben saber por qué se decide de cierta manera, lo que, a su vez, facilita la impugnación o cumplimiento de lo resuelto²⁹⁶. Si una decisión proviene de un sistema automatizado opaco, se dificultaría ese control. Por eso, los principios de **transparencia y explicabilidad** de la IA son esenciales para la seguridad jurídica. Instrumentos internacionales enfatizan este punto: la Carta Ética Europea sobre IA judicial – ya expuesta en líneas arriba - insiste en la **calidad y seguridad de los datos** y en la **transparencia, imparcialidad y equidad** de los algoritmos, permitiendo incluso auditorías externas para verificar su corrección.

Además, las resoluciones basadas en IA deben poder ser explicadas en lenguaje jurídico tradicional. En Costa Rica, para dar un ejemplo práctico, si el Poder Judicial usara un sistema de IA para, digamos, fijar la duración de penas o priorizar el trámite de ciertos asuntos, tendría que publicar los lineamientos de dicho sistema y garantizar que ante situaciones similares se obtengan resultados similares, salvo razón jurídica en contrario. De lo contrario, se quebrantaría la confianza pública y la coherencia del ordenamiento. En suma, la **seguridad jurídica demanda que la IA no introduzca arbitrariedad ni incertidumbre en la justicia**, sino que, por el contrario, se oriente a mejorar la consistencia y eficiencia del sistema respetando las garantías existentes.

➤ Derecho de Acceso a la Justicia

²⁹⁶ Ver al respecto, las siguientes sentencias: i. Sentencia 10461, Expediente 08-008818-0007-CO, Sala Constitucional; ii. Sentencia 00213, Expediente 13-000102-0640-CI, Sala Primera de la Corte; iii. Sentencia 00348, Expediente 10-000505-0164-CI, Tribunal Segundo Civil, Sección I; iv. Sentencia 00820, Expediente 00-100220-0468-CI, Tribunal Agrario; v. Sentencia 00067, Expediente 07-001446-0163-CA, Tribunal Contencioso Administrativo, Sección II; vi. Sentencia 05563, Expediente 11-004325-0007-CO, Sala Constitucional; vii. Sentencia 00027, Expediente 04-000665-0163-CA, Tribunal Contencioso Administrativo, Sección II.

El artículo 41 de la Constitución consagra el derecho de acceso a la justicia al señalar que “*ocurriendo a las leyes, todos han de encontrar reparación para las injurias o daños... Debe hacérseles justicia pronta, cumplida, sin denegación y en estricta conformidad con las leyes*”.

Este precepto le garantiza a toda persona la posibilidad real de acudir a los tribunales para reclamar sus derechos, obteniendo una resolución eficaz en un plazo razonable. La introducción de IA en la administración de justicia puede tener impactos, tanto positivos, como negativos, sobre este derecho. Por un lado, la tecnología bien aplicada puede **agilizar los procesos y reducir la mora judicial**, contribuyendo a la “justicia pronta y cumplida” que exige la Constitución. De hecho, el Poder Judicial costarricense ha buscado apoyarse en herramientas digitales para mejorar la celeridad: por ejemplo, ha implementado el **expediente electrónico** y plataformas en línea para interposición de recursos como medios de acercar la justicia a los usuarios. Un caso reciente es el plan piloto de IA en el Juzgado de Cobro Judicial de Pérez Zeledón, donde se desarrolló una nomenclatura inteligente que **clasifica automáticamente los escritos que ingresan y los agrupa por tema**, agilizando la gestión de la mayor cartera de casos (materia cobratoria)²⁹⁷. Tras dos años de uso, este sistema logró mayor celeridad procesal y mejor rendimiento en ese despacho, lo cual redunda en un acceso más rápido a la justicia para las partes involucradas. Estas innovaciones muestran el potencial de la IA para **mejorar la eficiencia judicial sin sacrificar garantías**, siempre que se apliquen con supervisión adecuada. Por otro lado, existe el riesgo de que la IA genere **nuevas barreras de acceso** si no se maneja con cuidado. La llamada “brecha digital” podría traducirse en una brecha de justicia: si ciertos usuarios (por edad, nivel educativo o ubicación geográfica) no logran interactuar con sistemas judiciales cada vez más digitalizados, su derecho de acceso se vería afectado. La Sala Constitucional ha reconocido el acceso a la justicia como un principio fundamental e incluso ha vinculado el acceso a Internet y a las tecnologías con el ejercicio efectivo de derechos fundamentales²⁹⁸.

²⁹⁷ Poder Judicial de Costa Rica, “Poder Judicial implementa inteligencia artificial para disminuir circulante en materia cobratoria”. Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/poder-judicial-implementa-inteligencia-artificial-para-disminuir-circulante-en-materia-cobratoria?catid=8&Itemid=409>

²⁹⁸ H. Miranda Bonilla, “El derecho de acceso a internet en la jurisprudencia de la sala constitucional de Costa Rica,” *Revista Jurídica Mario Alario D’Filippo* 13, no. 25 (2021): 5–18, <https://doi.org/10.32997/2256-2796-vol.13num.25-2021-3610>

Por ende, **la implementación de IA debe ir acompañada de políticas de inclusión digital**, asegurando que todos – especialmente poblaciones vulnerables – puedan beneficiarse de las mejoras tecnológicas. Además, el acceso a la justicia no solo implica entrar al sistema, sino también obtener un resultado comprensible. Si las resoluciones judiciales se vuelven excesivamente técnicas por la influencia de IA, podrían alejarse del ciudadano común. Es imperativo que, aunque se usen algoritmos complejos internamente, la cara visible para el usuario siga siendo humana y accesible. En conclusión, el derecho de acceso a la justicia demanda que la **IA sirva para acercar y no para alejar la justicia de las personas**, garantizando celeridad sin detrimento de la inclusión y la comprensión de las decisiones judiciales.

3.1.2.- Normativa Vigente Aplicable.

➤ **Protección de Datos Personales y Normativa Tecnológica**

La implementación de IA en la administración de justicia debe respetar estrictamente la **legislación de protección de datos personales**. En Costa Rica rige la *Ley N.º 8968 “Protección de la Persona frente al Tratamiento de sus Datos Personales”* (2011), que impone deberes a las instituciones en el manejo de datos sensibles. Un ejemplo de su interacción con la justicia es la referida adopción de una herramienta de IA para **despersonalizar sentencias judiciales**, eliminando datos sensibles antes de su publicación, en cumplimiento de la Ley 8968.

Esta normativa garantiza la privacidad de las partes en procesos judiciales, exigiendo que al publicar decisiones en plataformas como el sistema Nexus del Poder Judicial, se anonimicen nombres u otros datos personales. Cualquier sistema de IA que procese información de casos judiciales (expedientes, resoluciones, etc.) debe operar conforme a esta ley, asegurando **consentimiento o bases legales para el tratamiento de datos** y aplicando principios de minimización y seguridad. Además, Costa Rica cuenta con la *Ley de Certificados, Firmas Digitales y Documentos Electrónicos (N.º 8454)*, que proporciona validez jurídica a documentos y firmas digitales. Esta ley ha facilitado la digitalización de trámites judiciales (como notificaciones electrónicas y expedientes digitales) y sienta un precedente normativo para la adopción de tecnología en el Poder Judicial. Si bien no regula específicamente la IA, sí establece un entorno jurídico que **reconoce y valida el uso de medios tecnológicos** en actuaciones judiciales, siempre que se preserve la integridad y autenticidad de las actuaciones.

➤ Transparencia y Acceso a la Información Pública

La **transparencia en la función judicial** es otro pilar normativo relevante ante la introducción de IA. El artículo 30 de la Constitución garantiza el libre acceso a la información pública, principio desarrollado extensamente por la jurisprudencia de la Sala Constitucional²⁹⁹. Históricamente, aunque Costa Rica carecía de una ley específica de acceso a la información, la Sala IV tuteló este derecho vía recursos de amparo. Recientemente, en octubre de 2024, se aprobó la *Ley Marco de Acceso a la Información Pública* (N.º 10554), que obliga a todas las instituciones del Estado a **garantizar de forma proactiva, completa y oportuna el derecho de acceso a la información pública**.

Esta ley –ya en vigor tras su publicación en La Gaceta– refuerza la transparencia gubernamental y alcanza a los Poderes de la República, incluido el Judicial. En el contexto de la IA, dicha normativa implica que los ciudadanos podrían solicitar información sobre **sistemas algorítmicos usados por los tribunales**, su funcionamiento o criterios. La justicia abierta y la rendición de cuentas exigen que, si se emplean algoritmos para apoyar decisiones judiciales (por ejemplo, herramientas de apoyo en clasificación de casos o análisis jurisprudencial), exista **transparencia sobre su uso, alcance y limitaciones**. De hecho, el Poder Judicial costarricense impulsa iniciativas de *Justicia Abierta*³⁰⁰, publicando datos judiciales y estadísticas, lo cual deberá complementarse con información relativa a la IA para mantener la confianza pública. En resumen, las normas de transparencia actuales fomentan que la implementación de IA en la justicia se haga de forma abierta, permitiendo el escrutinio público y el acceso a la información pertinente, salvo en partes reservadas por ley (p. ej. datos personales o casos bajo confidencialidad).

➤ Ciberseguridad y Delitos Informáticos

Todo sistema de IA en la administración de justicia opera en entornos digitales, por lo que la **normativa de ciberseguridad** y delitos informáticos resulta aplicable. En 2013, mediante la

²⁹⁹ Ver al respecto los siguientes votos de la Sala Constitucional de la Corte Suprema de Justicia: Resolución n.º 598–1990, 30 de mayo de 1990; Resolución n.º 6240–1993, 26 de noviembre de 1993; Resolución n.º 136–2003, 15 de enero de 2003; Resolución n.º 10734–2004, 29 de septiembre de 2004; Resolución n.º 14519–2005, 21 de octubre de 2005; Resolución n.º 3454–2012, 9 de marzo de 2012; Resolución n.º 12046–2012, 1 de agosto de 2012.

³⁰⁰ Poder Judicial de Costa Rica, “Justicia Abierta”. Recuperado de: <https://servicios.poder-judicial.go.cr/index.php/funcionamiento-y-los-programas-pj/41-justicia-abierta>

Ley N.º 9135, Costa Rica reformó su legislación penal para tipificar diversos delitos informáticos, incorporando figuras como la violación de datos personales almacenados en redes informáticas, suplantación de identidad en el Internet, el espionaje de datos, etc. Estas disposiciones protegen las bases de datos y sistemas del Poder Judicial contra intrusiones o manipulaciones –un aspecto crítico si se utilizan algoritmos que manejan información sensible o ayudan en la toma de decisiones–. Además, tras incidentes de ciberataques de alto perfil en el país, se ha reconocido la necesidad de un marco más robusto. Actualmente, **se encuentra en desarrollo la “Ley de Ciberseguridad de Costa Rica”** (Expediente N.º 23.292)³⁰¹, iniciativa integral para regular la seguridad digital en el sector público y privado. Este proyecto, impulsado en 2023 con apoyo unánime en comisión, busca posicionar la ciberseguridad como política de Estado, obligando a instituciones a adoptar medidas de prevención, monitoreo y respuesta a incidentes.

Para la administración de justicia, la entrada en vigor de tal ley significaría estándares más altos de protección en cualquier plataforma tecnológica, incluyendo sistemas de IA. Asimismo, el Poder Judicial deberá seguir la *Estrategia Nacional de Ciberseguridad 2023-2027*³⁰², que establece protocolos para prevenir y responder a incidentes.

En materia de IA, esto se traduce en **asegurar la integridad, confidencialidad y disponibilidad** de los algoritmos y datos judiciales: por ejemplo, evitar que un sistema de IA sea alterado para sesgar sus resultados, o que datos judiciales usados para entrenarlo sean filtrados. En conclusión, la normativa vigente y emergente de ciberseguridad impone un marco de **deber de diligencia tecnológica** al Poder Judicial, que debe blindar sus herramientas digitales (incluyendo IA) contra usos malintencionados y cumplir con estándares éticos de seguridad.

3.1.3.- Políticas y Directrices Institucionales

La administración de justicia en Costa Rica se encuentra inmersa en un proceso de transformación digital acelerada, siguiendo una tendencia global que explora la integración de

³⁰¹ Asamblea Legislativa de la República de Costa Rica, *Proyecto de Ley – Ley de Ciberseguridad de Costa Rica*, Expediente N.º 23.292, presentado por José Joaquín Hernández Rojas y varios señores y señoritas diputados. Recuperado de: https://cicr.com/wp-content/uploads/2022/10/Exp_23292.pdf

³⁰² Sebastian May Grosser, “Gobierno presentó Estrategia Nacional de Ciberseguridad 2023-2027,” *Delfino CR*, 13 de noviembre de 2023. Recuperado de: <https://delfino.cr/2023/11/gobierno-presento-estrategia-nacional-de-ciberseguridad-2023-2027>

inteligencia artificial (IA) para mejorar la eficiencia sin comprometer principios jurídicos fundamentales. El Poder Judicial costarricense enfrenta desafíos crónicos como la mora judicial; por ejemplo, en la jurisdicción cobratoria (cobro judicial de deudas) se acumula **el 63.1 % de todos los casos pendientes**³⁰³.

Esta presión ha motivado la búsqueda de soluciones tecnológicas innovadoras. Paralelamente, el país ha desarrollado marcos institucionales y normativos para guiar el uso ético y responsable de la IA en el sector público, impulsados por entidades rectoras como el **Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones (MICITT)**. Este análisis examina exhaustivamente las **políticas, los planes y las directrices** pertinentes a la implementación de sistemas de IA en la justicia costarricense –desde las estrategias internas del Poder Judicial en digitalización y automatización, hasta las políticas nacionales de IA–, evaluando su impacto, viabilidad y los **desafíos y las oportunidades** que plantean en el ámbito jurisdiccional.

➤ Estrategias de Digitalización y Automatización en el Poder Judicial

El Poder Judicial de Costa Rica ha adoptado desde hace más de una década una estrategia progresiva de modernización tecnológica. En 2015 la Corte Suprema aprobó el **Plan Estratégico de Tecnologías de Información y Comunicación (PETIC 2015-2020)**, un instrumento orientado a la “*modernización de sus procesos, para un mejor aprovechamiento de la información en el quehacer institucional*”³⁰⁴. Este plan, construido de forma participativa, delineó **cuatro pilares estratégicos** –capacitación del recurso humano en TI, mejora de procesos internos de TI, orientación al usuario mediante servicios tecnológicos, y contribución al logro de objetivos judiciales– y sentó las bases para la transformación digital del Poder Judicial. La implementación del PETIC permitió importantes avances como la creación del **Expediente Judicial Electrónico** y la consolidación del programa “**Hacia Cero Papel**”, reduciendo, gradualmente, la dependencia del expediente físico en favor de sistemas digitales de gestión judicial.

³⁰³ Poder Judicial de Costa Rica, “Poder Judicial implementa inteligencia artificial para disminuir circulante en materia cobratoria”. Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/poder-judicial-implementa-inteligencia-artificial-para-disminuir-circulante-en-materia-cobratoria?catid=8&Itemid=409>

³⁰⁴ Poder Judicial de Costa Rica, “**Planificación estratégica dirigida a la modernización del Poder Judicial**.” Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/planificacion-estrategica-dirigida-a-la-modernizacion-del-poder-judicial?catid=8&Itemid=409#:~:text=Con%20el%20apoyo%20de%20la,informaci%C3%B3n%20en%20el%20que%20hacer%20institucional>

Con estos cimientos, el Poder Judicial entró a la década de 2020 con una infraestructura informática robusta –bajo la Dirección de Tecnología de la Información (DTI)– capaz de soportar nuevas soluciones. La pandemia de COVID-19 aceleró la adopción de plataformas electrónicas para garantizar la continuidad del servicio judicial, evidenciando la importancia de la tecnología. En 2021, por ejemplo, se contabilizaron más de **1.28 millones de expedientes judiciales electrónicos activos** y **2.36 millones de escritos presentados digitalmente**, indicadores de la masificación de los trámites en línea³⁰⁵. Autoridades judiciales reconocen que la innovación tecnológica es prioritaria para mantener la accesibilidad y eficiencia del sistema de justicia, así como para relacionarse de nuevas formas con las personas usuarias. Según Luis Guillermo Rivas, magistrado coordinador de la Comisión Gerencial de Tecnología, “*la tendencia de los Poderes Judiciales se dirige a establecer sus bases en plataformas informáticas y tecnologías innovadoras*”³⁰⁶ y el Poder Judicial costarricense ha comenzado a **incursionar en proyectos de IA** apoyado en su fortalecida infraestructura digital.

➤ Implementación de Sistemas de IA en la Judicatura

Sobre la base de esta transformación digital – como ya se ha expuesto superficialmente en acáptites anteriores - el Poder Judicial ha dado pasos iniciales en la incorporación de sistemas de IA para apoyar labores jurisdiccionales y administrativas. Uno de los casos pioneros es el comentado **plan piloto** desarrollado desde 2019 en el **Juzgado Especializado de Cobro de Pérez Zeledón**, donde se integró una herramienta de IA al sistema de gestión en línea para **tipificar y clasificar, automáticamente, los escritos que ingresan** al despacho. Mediante una nomenclatura entrenada por algoritmos, el sistema **selecciona y agrupa por tema los escritos** entrantes, agilizando su distribución interna. A dos años de iniciado el piloto, se reportaron mejoras significativas: “*mayor celeridad procesal, reducción de circulante y mejor rendimiento*” en dicho juzgado. El juez coordinador local subrayó que esta **IA de clasificación documental** permite una tramitación más ágil y rápida, contribuyendo a abatir la mora³⁰⁷. Dado el éxito, el Poder Judicial

³⁰⁵ Andrea Marín Mena, “**Poder Judicial avanza en la innovación de sus servicios a través de su fortalecimiento tecnológico**”, Poder Judicial de Costa Rica, 2022. Recuperado de: <https://cij.poder-judicial.go.cr/index.php/services/noticias/item/50-poder-judicial-avanza-en-la-innovacion-de-sus-servicios-a-traves-de-su-fortalecimiento-tecnologico#:~:text=,que%20atiende%20el%20Poder%20Judicial>

³⁰⁶ Ibid.

³⁰⁷ Poder Judicial de Costa Rica, “Poder Judicial implementa inteligencia artificial para disminuir circulante (,,)”.

anunció planes para **expandir este modelo de IA** a los 19 juzgados de cobro del país³⁰⁸, evidenciando confianza en la escalabilidad de la solución.

Otro frente de aplicación de IA ha sido la **protección de datos personales en sentencias**. En cumplimiento de la Ley N° 8968 “Protección de la Persona frente al Tratamiento de sus Datos Personales”, el Poder Judicial –a través de su Comisión de Protección de Datos– implementó en 2022 una “*herramienta para la despersonalización de sentencias judiciales*”, apoyada en **mecanismos de IA**³⁰⁹. Esta herramienta automatiza el proceso de **anonimización de datos sensibles** en las resoluciones publicadas en el sistema Nexus, removiendo nombres y detalles privados de las partes, de manera ágil y conforme con los mandatos legales de privacidad. La iniciativa garantiza la tutela de derechos de las personas usuarias (protección de datos personales) sin sacrificar la transparencia del acervo jurisprudencial. Cabe destacar que la institución complementó la innovación tecnológica con **acciones normativas internas** para hacer cumplir dicho mandato, mostrando un enfoque integral donde la regulación y la tecnología van de la mano.

En el ámbito de la investigación penal, la Policía Judicial (OIJ), órgano auxiliar del Poder Judicial, también apuesta por la IA para potenciar la eficacia investigativa. Actualmente desarrolla un proyecto ambicioso denominado “**SUPERCOP**” (*Sistema Único Policial Especializado en la Resolución de la Criminalidad*), un sistema de automatización con capacidad de **autoaprendizaje** basado en técnicas de *machine learning*. Según el director del OIJ, la IA permitirá “*automatizar procesos y ser mucho más rápidos en la resolución de casos*”, acortando tiempos en pesquisas y abordando análisis que hoy consumirían ingentes recursos humanos³¹⁰. El sistema SUPERCOP procesará **millones de datos, archivos, imágenes, videos y audios**, aplicando algoritmos para encontrar patrones y correlaciones útiles en la investigación criminal. Se proyecta incluso la **predicción de eventos delictivos y apoyo a la toma de decisiones** policiales, reduciendo los

³⁰⁸ Poder Judicial de Costa Rica, “Juzgados Especializados de Cobro de San José se preparan para trabajar con Inteligencia Artificial,” 7 de junio de 2023. Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/juzgados-especializados-de-cobro-de-san-jose-se-preparan-para-trabajar-con-inteligencia-artificial?catid=8&Itemid=409#:~:text=Juzgados%20Especializados%20de%20Cobro%20de,a%20ponerse%20en%20marcha>

³⁰⁹ Poder Judicial de Costa Rica, “Novedosa herramienta de Inteligencia Artificial se aplica en mejora de la protección de datos (...”)

³¹⁰ Marisel Rodríguez Solís, "OIJ apuesta por la inteligencia artificial," Poder Judicial, 23 de enero de 2023. Recuperado de: <https://pjenlinea3.poder-judicial.go.cr/biblioteca/uploads/Archivos/Articulo/OIJ%20apuesta%20por%20la%20inteligencia%20artificial.pdf>

sesgos conscientes o inconscientes presentes en el factor humano. Si bien estas expectativas deben tomarse con cautela –la promesa de eliminar sesgos mediante IA exige garantizar la calidad y objetividad de los datos de entrenamiento–, el proyecto refleja la visión institucional de la IA como un “aliado poderoso” para **reforzar la labor policial** y, en última instancia, mejorar la administración de justicia penal.

En suma, el Poder Judicial costarricense ha comenzado a **integrar sistemas de IA** en tareas específicas: la gestión de expedientes (clasificación de escritos), la difusión de jurisprudencia (anonimización de sentencias) y la investigación criminal (análisis inteligente de grandes datos). Estas experiencias piloto demuestran voluntad de innovación y ofrecen primeras evidencias de impacto positivo (mayor rapidez, eficiencia y cumplimiento normativo). No obstante, su alcance aún es limitado y se desarrollan bajo marcos experimentales; de ahí la importancia de examinar las **directrices institucionales y políticas** que orientan y delimitan el uso de la IA en el entorno judicial.

➤ **Directrices Institucionales y Marco Normativo en el Ámbito Judicial**

Dentro del Poder Judicial, órganos como la **Comisión de Protección de Datos** y la **Comisión Gerencial de Tecnología de la Información** actúan como garantes y promotores de buenas prácticas en el manejo de información y TI. Si bien hasta la fecha no han trascendido al público lineamientos específicos sobre IA emitidos por la Corte Suprema o el Consejo Superior, es claro que las máximas autoridades judiciales están conscientes de la relevancia y riesgos de estas tecnologías. En el *Foro “TICs en la Administración de Justicia: Retos y oportunidades post pandemia”* (2022), el magistrado Rivas Loáiciga destacó que la justicia debe “*adaptarse y adoptar*” el uso de tecnologías en sus decisiones, **sin olvidar principios como la imparcialidad, igualdad y trato equitativo** a las partes³¹¹. Esta declaración subraya un criterio orientador: la incorporación de IA no puede menoscabar las garantías procesales ni la igualdad de armas, valores esenciales del debido proceso. Del mismo modo, la discusión internacional reflejada en ese foro

³¹¹ Poder Judicial de Costa Rica, “*Analizan los retos y oportunidades post pandemia de las Tics en la Administración de Justicia*,”. Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/analizan-los-retos-y-oportunidades-post-pandemia-de-las-tics-en-la-administracion-de-justicia?catid=8&Itemid=409#:~:text=%E2%80%9CSe%20destaca%20que%20la%20inteligencia,Uni%C3%B3n%20Europea%E2%80%9D%2C%20detall%C3%A9n%20Bujosa%20Vadell>

recalcó consideraciones éticas: “*la inteligencia artificial no es un fin en sí mismo, sino un medio que debe servir a las personas..., su fiabilidad debe estar garantizada y centrada en el ser humano*”³¹², en palabras de un experto del Centro Nacional de Tribunales de EE.UU., enfatizando que el marco normativo en construcción busca reforzar **la confianza, la dignidad humana, la libertad, la democracia, el Estado de Derecho y los derechos humanos**. Este enfoque coincide con la perspectiva que inspira los esfuerzos costarricenses: la IA judicial debe ser **antropocéntrica**, confiable y sujeta a control humano permanente.

En términos de **directrices institucionales**, el Poder Judicial ha emitido políticas generales que, si bien no abordan la IA de forma explícita, crean un entorno propicio para su aplicación. La política de “**Justicia Abierta**” fomenta la transparencia y la publicación de datos judiciales, lo cual puede considerarse un habilitador para proyectos de IA (al proveer datos abiertos para entrenamiento de modelos, siempre que se resguarden los datos personales). Asimismo, programas como “**Un Mejor Poder Judicial**” y “**Buenas Prácticas**” promueven la innovación y mejora continua en la gestión judicial, legitimando internamente la experimentación con nuevas herramientas tecnológicas. Incluso el esfuerzo “*Hacia Cero Papel*” –orientado a eliminar el papel en los procesos– implica la **digitalización de expedientes y documentos** a gran escala, un prerequisito para la aplicación de algoritmos avanzados de búsqueda, análisis predictivo o minería de textos en los archivos judiciales. Cada una de estas iniciativas configura componentes de un **ecosistema digital** dentro del cual la IA puede integrarse gradualmente, siempre que se armonice con los principios jurídicos y lineamientos éticos mencionados.

Por otra parte, es importante mencionar que la **Corte Plena** (el órgano colegiado superior del Poder Judicial) participa activamente en la discusión legislativa nacional sobre IA. En noviembre de 2023³¹³, por ejemplo, en su sesión ordinaria se conoció y comentó el *Proyecto de Ley “Ley para la Promoción Responsable de la Inteligencia Artificial en Costa Rica”* (expediente N° 23.919).

³¹² Ibid.

³¹³ Poder Judicial de Costa Rica, “Agenda de Corte Plena – Lunes 6 de noviembre de 2023,” 6 de noviembre de 2023. Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/agenda-de-corte-plena-lunes-6-de-noviembre-de-2023?catid=8&Itemid=409#:~:text=para%20la%20promoci%C3%B3n%20responsables%20de,919>

Vale indicar aquí que la judicatura, como poder independiente del Estado, suele emitir **criterios técnicos o constitucionales** sobre iniciativas de ley que puedan impactar la administración de justicia. Su involucramiento temprano en este debate sugiere interés en asegurar que cualquier normativa futura sobre IA sea **compatible con las necesidades del quehacer judicial** y respete la independencia y competencias propias del Poder Judicial. En resumen, aunque todavía no exista un reglamento interno específico sobre IA, el **marco normativo vigente –leyes de protección de datos, principios constitucionales, políticas institucionales de transparencia y eficiencia– y la conciencia expresada por magistrados y comisiones**– funciona ya como guía y condicionante para la implementación de sistemas inteligentes en tribunales.

➤ **Políticas Nacionales de IA en la Administración Pública Costarricense**

A nivel país, Costa Rica ha dado pasos firmes en el establecimiento de **políticas públicas de inteligencia artificial** en los últimos años, reconociendo el potencial transformador de esta tecnología en todas las áreas del Estado, incluida la justicia. El MICITT, como ente rector en innovación y tecnología, lanzó en octubre de 2024 la **Estrategia Nacional de Inteligencia Artificial (ENIA) 2024-2027³¹⁴**, la primera política pública costarricense dedicada enteramente a la IA. Con ello, Costa Rica se convirtió en el **primer país de Centroamérica** en contar con una estrategia nacional orientada a guiar el uso, la adopción y el desarrollo de la IA de manera íntegra³¹⁵.

La ENIA establece un marco detallado para fomentar el **uso responsable y ético** de la IA, promoviendo principios fundamentales como la **dignidad humana, la supervisión humana, la transparencia, la equidad, la inclusión social y el desarrollo sostenible**. Estos principios buscan asegurar que el desarrollo de la IA esté **centrado en las personas** y orientado al bienestar social, evitando agravar brechas de desigualdad. Se trata de una visión alineada con los valores democráticos y de derechos humanos, en sintonía con las recomendaciones de organismos

³¹⁴ Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones (MICITT), Estrategia Nacional de Inteligencia Artificial de Costa Rica (San José, C.R.: MICITT, 2024), ISBN 978-9968-732-94-9. Recuperado de: <https://www.micitt.go.cr/sites/default/files/2024-10/Estrategia%20Nacional%20de%20Inteligencia%20Artificial%20de%20Costa%20Rica%20ESP.pdf>

³¹⁵ Samantha Brenes Mora, “Micitt lanza primera política pública sobre uso y desarrollo de Inteligencia Artificial,” *Delfino.cr*, 24 de octubre de 2024, 3:54 p. m. Recuperado de: <https://delfino.cr/2024/10/micitt-lanza-primera-politica-publica-sobre-uso-y-desarrollo-de-inteligencia-artificial>

internacionales y experiencias comparadas (Chile, Brasil, Uruguay, entre otros, ya contaban con estrategias similares)³¹⁶.

Un aspecto medular de la ENIA es que contempla la creación de un **marco normativo específico** para IA, que establezca pautas técnicas y éticas claras para su desarrollo y uso en el país (pág. 56). En otras palabras, anticipa la necesidad de legislación o regulaciones sectoriales que traduzcan esos principios en obligaciones concretas para instituciones y empresas. Asimismo, la estrategia prioriza el fortalecimiento de la infraestructura digital habilitadora: destaca la importancia del **despliegue de redes 5G** (para soportar aplicaciones intensivas de datos e IA en tiempo real) y propone la creación de un **Centro Nacional de Excelencia en IA**. Este centro serviría como *hub* para la investigación, capacitación de talento humano y desarrollo de soluciones de IA en el contexto local. Desde la perspectiva de la administración de justicia, tales medidas podrían traducirse en mayor disponibilidad de herramientas y conocimientos especializados que las instituciones judiciales podrían aprovechar; por ejemplo, el Centro de Excelencia podría colaborar con el Poder Judicial en proyectos piloto, compartir mejores prácticas o proveer asesoría en la adopción de IA con estándares de calidad y ética.

Autoridades del MICITT han señalado que la IA “*tiene el potencial de revolucionar la forma en que brindamos servicios a la ciudadanía, haciendo al Estado más eficiente, más transparente y más cercano a las personas*”³¹⁷. Esta visión de un “gobierno inteligente” basado en datos y evidencia ciertamente incluye al sector Justicia, que es un proveedor crítico de servicios públicos. De hecho, entre las primeras iniciativas concretas estuvo la participación de Costa Rica en **fAIR LAC** (Fairness, Accountability, Innovation and Responsible AI in LAC), una iniciativa del Banco Interamericano de Desarrollo para impulsar el uso ético de la IA. Ya en 2021 se lanzó *fAIR LAC Costa Rica* con la colaboración de múltiples actores públicos (MICITT, Ministerio de Educación, Caja de Seguro Social, etc.) y privados, con el objetivo de “**promover, educar y regular el**

³¹⁶ Álvaro Murillo, “Gobierno lanza estrategia nacional sobre Inteligencia Artificial,” *Semanario Universidad*, 24 de octubre de 2024. Recuperado de: <https://semanariouniversidad.com/pais/gobierno-lanza-estrategia-nacional-sobre-inteligencia-artificial/#:~:text=La%20ENIA%20procura%20pautas%20t%C3%A9cnicas,materia%3A%20Chile%2C%20Brasil%20y%20Uruguay>

³¹⁷ Samantha Brenes Mora, “Micitt Lanza primera política pública”.

desarrollo y uso ético de la IA" mediante casos piloto y estándares de referencia³¹⁸. Esta alianza multisectorial –en la que el Poder Judicial no figura directamente, pero cuyos lineamientos pueden influir en todo el aparato estatal– refuerza la idea de que Costa Rica busca **ser pionera en la adopción responsable de IA**, anticipando riesgos éticos y creando capacidad institucional para enfrentarlos.

Paralelamente a la ENIA, en la Asamblea Legislativa han surgido proyectos de ley para dotar de sustento jurídico a la política de IA:

- **Proyecto de Ley N.º 23.771: Ley de Regulación de la Inteligencia Artificial**

Presentado el 30 de mayo de 2023, fue uno de los primeros esfuerzos integrales por regular la IA en Costa Rica. Este proyecto (elaborado con asistencia de IA según su exposición de motivos) propone la creación de un marco jurídico para el *desarrollo, implementación y uso de la IA*, en consonancia con los derechos y principios constitucionales.

En primer lugar, debe decirse que la metodología empleada en la elaboración del proyecto de "Ley de Regulación de la Inteligencia Artificial en Costa Rica" mediante la utilización de ChatGPT-4 a partir de un único *prompt* constituye una seria deficiencia en términos de técnica legislativa y gobernanza democrática que compromete la legitimidad y calidad técnica del instrumento normativo. En primer término, desde una perspectiva de técnica legislativa, la elaboración de un cuerpo normativo de tal envergadura y complejidad requiere un proceso metodológico riguroso que involucre, como mínimo: (i) un análisis exhaustivo del marco jurídico existente y su interacción con la norma propuesta; (ii) un estudio comparado de experiencias regulatorias en otras jurisdicciones; (iii) una evaluación de impacto regulatorio que examine las consecuencias previsibles de la normativa; y (iv) un proceso de consulta con expertos sectoriales y stakeholders relevantes. La pretensión de sustituir este proceso metodológico por un único

³¹⁸ Banco Interamericano de Desarrollo (BID), "Costa Rica promoverá el uso responsable de la inteligencia artificial con apoyo del BID," comunicado de prensa, San José, 29 de septiembre de 2021. Recuperado de: <https://www.iadb.org/es/noticias/costa-rica-promovera-el-uso-responsable-de-la-inteligencia-artificial-con-apoyo-del-bid#:~:text=San%20Jos%C3%A9%2C%2029%20de%20septiembre,Instituto%20Tecnol%C3%B3gico%20de%20Costa>

prompt a un modelo de lenguaje, por avanzado que este sea, constituye una simplificación inadmisible del proceso legislativo.

Más aún, resulta particularmente problemático que un proyecto destinado a regular la inteligencia artificial haya sido elaborado precisamente mediante una implementación acrítica y metodológicamente deficiente de esta tecnología. Esta circunstancia no solo compromete la calidad técnica del proyecto - que inevitablemente refleja las limitaciones del modelo de lenguaje utilizado - sino que, también, socava su credibilidad como instrumento regulatorio.

La complejidad de regular la inteligencia artificial exige un proceso legislativo que combine el conocimiento técnico especializado con una amplia participación social y una rigurosa evaluación de impacto. Si bien las herramientas de IA pueden ciertamente contribuir a este proceso - por ejemplo, en el análisis de legislación comparada o en la identificación de potenciales impactos regulatorios - su utilización debe enmarcarse en una metodología más amplia que garantice la calidad técnica y la legitimidad democrática del producto normativo.

Ahora bien, más allá de esta disquisición crítica inicial, el articulado se compone de ocho capítulos que, en conjunto, configuran un diseño institucional y axiológico exhaustivo, abarcando desde la declaratoria de principios rectores de la IA hasta la instauración de mecanismos de evaluación y auditoría. La lectura detenida del articulado revela una intención manifiesta de incorporar no solo la Constitución Política de 1949 como piedra de toque, sino, también, de invocar diversas experiencias comparadas, especialmente las protagonizadas por la Unión Europea y las recomendaciones internacionales sobre ética de la IA.

El texto inicia con un **Capítulo I**, titulado Disposiciones Generales. El artículo 1, al delinear el objeto de la ley, recalca la necesidad de garantizar que la IA se desarrolle, implemente y utilice “en concordancia con los principios y derechos establecidos en la Constitución Política de 1949”. Resulta encomiable la vocación de anclar la IA en la dignidad humana, la protección de la persona y la observancia de valores democráticos, a tenor de lo que prescribe la jurisprudencia constitucional costarricense. Su redacción, por otra parte, pone de relieve la impronta del asistente virtual (ChatGPT-4), en tanto se advierte un estilo directo que, si bien claro, carece en ocasiones

de la precisión jurídica característica de los proyectos de ley redactados por asesores con experiencia prolongada.

El artículo 2 ofrece definiciones fundamentales, entre las cuales destacan las de “Inteligencia Artificial (IA)”, “Agente de Inteligencia Artificial” y “Sesgo algorítmico”. Sorprende la amplitud conceptual de IA, descrita como un “conjunto de tecnologías y algoritmos” que permite a los sistemas “realizar tareas y tomar decisiones de manera autónoma, imitando las capacidades humanas”. Se advierte un hilo conductor con la doctrina que asocia la IA al razonamiento y el aprendizaje emulado, si bien el artículo no especifica categorías de riesgo ni se adentra en distinciones sobre aplicaciones de alto o bajo impacto.

La ley enuncia un catálogo de principios en su **Capítulo II** (artículos 3 y 4). Se subrayan la Equidad, la Responsabilidad, la Transparencia y la Seguridad. El principio de responsabilidad, en particular, enfatiza la obligación de “asumir la responsabilidad de sus acciones y decisiones, minimizando los riesgos y garantizando el cumplimiento de los principios éticos y legales”. Esta referencia implícita a la *accountability* algorítmica concuerda con la tendencia internacional, mas se echa de menos una regulación pormenorizada que determine cómo, cuándo y ante qué autoridad se ejercerá dicha rendición de cuentas. El artículo 4, por su parte, encuadra la aplicación de la Constitución Política, destacando la dignidad humana y la igualdad. Esto retoma la tutela de derechos como el de la vida privada (art. 24) o la libertad de expresión (art. 28), con miras a recalcar que cualquier despliegue de IA que comprometa datos o incida en la formación del debate ciudadano debe regirse por los cánones constitucionales de proporcionalidad y debido proceso.

La sección concerniente al registro de sistemas de IA, comprendida en el **Capítulo III** (artículos 5 y 6), es uno de los rasgos distintivos de este proyecto, pues impone a los desarrolladores la carga de registrar sus sistemas ante “la autoridad competente”, con un elenco de información relativa a algoritmos, medidas de seguridad y objetivos de la tecnología. La amplitud de dicha obligación podría suscitar interrogantes acerca de la viabilidad y eficacia del registro; en particular, sería preciso especificar criterios de clasificación o segmentación de los sistemas susceptibles de registro. El artículo 6, que regula supervisión y auditoría a través de la “autoridad competente”, no define con claridad cuál es ese ente ni si se trata de un nuevo organismo o de una dependencia existente. Tal omisión, previsible tal vez en un texto con génesis eminentemente

experimental, deberá ser resuelta en la elaboración del reglamento, a fin de evitar solapamientos institucionales o vacíos de competencia.

En el **Capítulo IV** (artículos 7 al 10), la iniciativa da un paso hacia la especificidad técnica. El artículo 7 incorpora la Evaluación de Impacto para sistemas de “alto riesgo”, inspirada aparentemente en los marcos regulatorios europeos, a fin de mitigar sesgos, reforzar la explicabilidad y calibrar la proporcionalidad de la IA en ámbitos sensibles. No obstante, se echa de menos una tipificación clara de lo que se entenderá por “alto riesgo”; la norma se limita a una formula genérica, delegando la concreción en una posterior reglamentación. El artículo 8, sobre Sesgos y Discriminación, se adentra en la aludida problemática del “sesgo algorítmico”. Reitera la necesidad de utilizar “conjuntos de datos representativos y diversificados” y de establecer revisiones periódicas de algoritmos. Si bien la intención es loable, la ley podría encontrarse en la necesidad de integrar pautas más robustas de auditoría, so pena de que la exigencia se diluya en voluntarismos. El artículo 9, por su parte, clama por la “explicabilidad” de los sistemas de IA, reiterando que se proporcione información comprensible sobre las razones y criterios que subyacen a las decisiones algorítmicas. Esta directriz confluye con la tendencia jurídica internacional a exigir “razonabilidad algorítmica”. Pero, nuevamente, la ausencia de un procedimiento formal o de un estándar técnico de explicabilidad deja un margen de interpretación.

La protección de datos personales (artículo 10) reafirma la necesidad de enmarcar el tratamiento en las disposiciones constitucionales y en la ley costarricense de datos (Ley 8968), incorporando la exigencia de consentimiento informado y el derecho de acceso y rectificación. Con ello, se conjugan los principios de autodeterminación informativa con la arquitectura conceptual de la IA, pretendiendo asegurar que la manipulación intensiva de datos masivos no atropelle la esfera íntima de los titulares.

En el **Capítulo V** (artículo 11) se enumeran diversos campos sectoriales (salud, finanzas, transporte, educación, justicia, administración pública) donde la implementación de IA quedaría sujeta a regulación especial. Esta decisión de sectorizar la IA podría tener repercusiones positivas, al permitir que cada ámbito desarrolle reglas específicas según su idiosincrasia. Sin embargo, la redacción sumaria del artículo no aclara si el legislador pretende exigir leyes sectoriales ulteriores o si bastará con directrices de la misma “autoridad competente”. El capítulo VI (artículos 12 y 13)

prevé la creación de la Autoridad Reguladora de Inteligencia Artificial (ARIA) e introduce un régimen de sanciones de carácter proporcional. La sola mención de ARIA, sin detalles sobre su composición, independencia o recursos, deja en el aire la forma de su instauración. Se infiere que su diseño quedaría igualmente en manos del reglamento o de normas complementarias, con la consiguiente dependencia de la voluntad política.

El artículo 14 concierne la relación con la Ley General de Telecomunicaciones, aludiendo a la gestión del espectro radioeléctrico, la protección de la privacidad en la prestación de servicios y la seguridad de redes donde convergen sistemas de IA. Ello se corresponde con la creciente importancia de la IA en las infraestructuras críticas de telecomunicaciones, pero no se profundiza en la interrelación con la Superintendencia de Telecomunicaciones (SUTEL) o con eventuales potestades concurrentes de la ARIA. Con posterioridad, el artículo 15 retoma la protección de derechos humanos: no discriminación, privacidad de datos y la garantía de explicabilidad. La referencia expresa al acceso a la justicia, al derecho al trabajo y a la formación y reubicación de trabajadores afectados por automatizaciones de IA sugiere la intención de ofrecer un paraguas de garantías comprehensivo.

El artículo 16 introduce la figura del “Informe de Impacto en Derechos Humanos”. Resulta interesante que se exija a los desarrolladores y usuarios de sistemas de IA la elaboración de un documento previo que evalúe riesgos y efectos negativos en derechos fundamentales. Este tipo de obligación, reminiscencia de las “Data Protection Impact Assessments”³¹⁹ del RGPD europeo, asume la posibilidad de que la IA incida en derechos consagrados en la Constitución. La complejidad de este informe, sin embargo, obligaría a la emisión de guías y estándares concretos. Por su parte, el artículo 17, referente a la capacitación y concienciación, fomenta la formación de profesionales en ética y derechos humanos asociados al uso de la IA. De nuevo, ello denota una mirada transversal, si bien las obligaciones específicas de la administración para generar tal capacitación se relegan a un ámbito más programático.

³¹⁹ **Agencia Española de Protección de Datos**, “Data Protection Impact Assessments (Evaluación de Impacto relativa a la Protección de Datos),” en *Portal de la Agencia Española de Protección de Datos*, última actualización el 16 de octubre de 2024. Recuperado de: <https://www.aepd.es/en/rights-and-duties/fulfill-your-duties/measures-compliance/data-protection-impact-assessments#:~:text=El%20RGPD%20introduce%20el%20concepto,que%20s%C3%AD%20requieren%20su%20realizaci%C3%B3n>

El Capítulo VIII, ya en las disposiciones finales (artículos 18 y 19), acentúa la naturaleza escalonada de la implementación y necesidad de actualizar la ley a la luz de la evolución tecnológica. Se compromete, por ende, a una evaluación periódica y a una posibilidad de modificación flexible. Este rasgo adaptativo es coherente con la volatilidad propia de la IA, aunque quizá sea menester un protocolo más definido de revisión legislativa (consultas públicas, informes obligatorios del ente regulador), para que la lógica adaptativa no quede sujeta al arbitrio político coyuntural.

En conclusión, la “Ley de Regulación de la Inteligencia Artificial en Costa Rica (2023)” aspira, en su estructura, a forjar un marco multidimensional que combine la responsabilidad algorítmica, la protección de datos, la supervisión institucional y la observancia de los más altos postulados constitucionales. La confluencia de disposiciones generales y sectoriales, sumada a la pretensión de crear un ente regulador autónomo (ARIA) y de imponer la obligación de evaluaciones de impacto, sugiere un influjo evidente de las tendencias regulatorias de la Unión Europea y otras jurisdicciones con larga tradición de *cautionary approach* en IA.

Resulta innegable, no obstante, que el proyecto expone cierta laxitud en la precisión técnica y en la definición del armazón institucional, probablemente consecuencia de su génesis a partir de la asistencia de ChatGPT-4. Aunque la dimensión experimental de esta metodología de redacción no anula en su totalidad la validez intrínseca del texto, aboca a que, en la tramitación parlamentaria, se aborde con meticulosidad la explicitación de normas de jerarquía secundaria, la identificación detallada de competencias y la graduación de sanciones. Asimismo, sería aconsejable una refinación conceptual en torno a la clasificación de riesgo, la distinción de funciones entre la ARIA y otras autoridades sectoriales y la instauración de un régimen de auditoría algorítmica robusto.

En lo que atañe a los derechos humanos, el proyecto converge oportunamente con la doctrina que insiste en la dignidad humana y la no discriminación, introduciendo obligaciones sustantivas de transparencia y explicabilidad. Sin embargo, la eficacia real de tales preceptos se verá supeditada a la existencia de instrumentos concretos de verificación y sanción, así como a la disposición de las instituciones públicas de constituir unidades técnicas capaces de escrutar la opacidad algorítmica. La pertinencia de un capítulo específico para la IA en el ámbito judicial —

considerando la altísima sensibilidad de la toma de decisiones automatizadas en procesos jurisdiccionales— se echa de menos, aun cuando el artículo 11(e) hace un esfuerzo al sector justicia.

En suma, esta iniciativa exhibe la voluntad de diseñar una normativa comprehensiva que, sin rehuir la innovación, la someta a criterios de equidad, seguridad y responsabilidad. Cabe esperar que, en su posterior análisis y enriquecimiento por parte del foro legislativo, se depuren las lagunas existentes, se delimiten con mayor finura los alcances de las obligaciones de registro y supervisión y se perfeccione el régimen sancionatorio.

- **Proyecto de Ley N.º 23.919: Ley para la Promoción Responsable de la IA en Costa Rica**

El proyecto de ley contenido en el expediente número 23.919, titulado “Ley para la Promoción Responsable de la Inteligencia Artificial en Costa Rica”, se configura como una tentativa de instituir un marco jurídico que, en términos de alcance y lineamientos, procura conciliar el ímpetu innovador propio de la inteligencia artificial con la exigencia ineludible de salvaguardar valores jurídicos y principios éticos de raigambre constitucional. Si bien su exposición de motivos insiste en el potencial transformador de la inteligencia artificial para impulsar el desarrollo económico, social y científico del país, su articulado revela, al mismo tiempo, un notable acento en la protección de derechos fundamentales, la gobernanza ética y la transparencia en la adopción de dichas tecnologías.

El artículo 1 puntualiza con especial énfasis que la ley tiene por objeto promover, de forma responsable y ética, la investigación, el uso y despliegue de la inteligencia artificial, considerando la dignidad humana y la transparencia como ejes fundamentales. Desde la propia noción de “promoción responsable” se advierte la pretensión de alentar iniciativas tecnológicas en el ámbito de la IA, sin, por ello, sacrificar los valores esenciales de la persona y la prudente supervisión estatal que se infiere de diversas normas constitucionales y tratados internacionales. Aunado con lo anterior, el artículo 2 declara la inteligencia artificial como de interés público, lo que confirma la intención de subrayar su carácter transversal y estratégico para la competitividad del país y la eficiencia del sector público.

Conforme se avanza en el articulado, se advierte que el artículo 3 extiende su ámbito de aplicación a toda persona física o jurídica que, en el territorio nacional, investigue, desarrolle, aplique o utilice sistemas basados en IA, fijando así un radio de acción amplio. La ausencia de categorías específicas de riesgo en esta sección podría, sin embargo, plantear dudas sobre la proporcionalidad de algunas obligaciones a proyectos de distinto calado o complejidad. En el marco de la regulación comparada, por ejemplo, la Unión Europea ha optado por distinguir niveles de riesgo y exigir cautelas diversas según la intensidad del posible impacto de la IA en derechos y bienes jurídicos relevantes. Este proyecto costarricense, en cambio, presenta un diseño más generalista que, aunque de lectura sencilla, podría requerir posterior reglamentación para discernir entre aplicaciones de bajo riesgo y sistemas de riesgo alto.

En el artículo 5 se enumeran los principios rectores de la nueva regulación, destacando la confiabilidad, la supervisión humana, la privacidad, la equidad y la no discriminación, la protección del medio ambiente y la proporcionalidad. El acento sobre la supervisión humana, que en la práctica se materializa en la exigencia de que exista un control efectivo de la persona en la toma de decisiones, resulta consecuente con la tradición jurídica costarricense, que reserva a los operadores humanos la valoración última cuando se vean comprometidas garantías como el derecho de defensa, el debido proceso o la intimidad. Aun cuando las menciones a la transparencia y a la explicabilidad algorítmica se exhiben en dicho artículo 5 de manera algo genérica, su sola incorporación expresa denota una aproximación a los estándares éticos postulados por organismos internacionales referidos en este trabajo de investigación. No obstante, cabría profundizar en el modo en que se exigirá y verificará la explicabilidad de algoritmos complejos, so pena de que el principio se diluya en la abstracción.

El artículo 6 asigna la rectoría de la IA al Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones, anunciando la necesidad de una política pública coherente. Sin embargo, es el artículo 7, al crear la Comisión Interinstitucional para el Desarrollo de la Inteligencia Artificial, el que dispone la verdadera columna vertebral de la gobernanza prevista por el proyecto. Este órgano, adscrito al MICITT, se compondrá de distintas carteras ministeriales, así como de representantes del sector privado y académico, con el cometido de conocer y dar visto bueno a los proyectos, dictar planes y proponer reglamentos que operativicen la promoción y control de la IA. Suerte de dirección colegiada, la Comisión podrá evaluar iniciativas declaradas de interés público,

fomentar alianzas público-privadas y promover la capacitación de los funcionarios, lo que se advierte como un avance en la institucionalidad costarricense en materia de innovación tecnológica. La amplitud de sus competencias, no obstante, requeriría una reglamentación detallada que delimite la interacción con otros entes como la Agencia de Protección de Datos de los Habitantes (PRODHAB), sobre todo cuando se trate de proyectos masivos o que impliquen uso intensivo de datos personales.

El artículo 9 añade la figura del comité ético, técnico y científico, al que incumbe asesorar a la Comisión en la valoración de proyectos con implicaciones de alto riesgo. Su función consiste en examinar y recomendar las propuestas a la luz de principios éticos y de salvaguarda de derechos fundamentales. Esta dimensión se antoja pertinente en un contexto tecnológico que a menudo suscita riesgos de discriminación algorítmica, vulneraciones a la privacidad o afectaciones al trabajo humano. Sin embargo, el texto no pormenoriza la metodología de revisión ni el procedimiento para la emisión de dictámenes, lo cual podría exigir un desarrollo reglamentario para que el referido comité realice, por ejemplo, evaluaciones de impacto de manera estandarizada.

La voluntad de facilitar la innovación se plasma con fuerza en el artículo 30, donde se contempla la posibilidad de crear “espacios controlados de prueba” o sandboxes regulatorios, con objeto de permitir la experimentación en entornos de riesgo acotado. Este acercamiento cohonesta la dimensión de “promoción” anunciada en el título de la ley, al propiciar que empresas emergentes, PYMES o entidades públicas puedan desarrollar prototipos de inteligencia artificial, bajo la supervisión de la Comisión y con un régimen de temporalidad, para constatar su eficacia y riesgos antes de ser escalados a mayor escala. Dicha medida se complementa con la previsión de prototipos de regulación en el artículo 31, una modalidad que calca la práctica de la innovación regulatoria adoptada en países avanzados.

En materia de derechos y salvaguardas, el proyecto dedica diversas disposiciones a la protección de los datos personales, como el artículo 13 y el artículo 15, reafirmando la vigencia de la Ley 8968, además de subrayar la obligación de obtener consentimiento informado para tratamientos que involucren datos biométricos. Este acierto normativo refuerza los lineamientos de la Sala Constitucional en materia de autodeterminación informativa, aunque se echa de menos un capítulo más sólido que describa, con nitidez, en qué términos la Comisión o el Comité Ético

verificarán la adecuada protección de datos cuando se trate de sistemas de IA que abarquen grandes volúmenes de información sensible.

La responsabilización de los desarrolladores y usuarios de IA, prevista en el artículo 32, sigue el hilo de la responsabilidad objetiva y solidaria, si bien el texto no detalla la concreción del régimen sancionatorio ni los criterios técnicos que permitan graduar sanciones según el tipo o alcance de la infracción. Ello contrasta con la minuciosidad que, en otras jurisdicciones, se dedica a la diferenciación de infracciones leves, graves y muy graves, y a la definición de un correlato sancionador. En consecuencia, el proyecto establece una obligación genérica de compensar daños, pero deja en el aire la relación directa con la normativa civil y con la posible intervención penal si se manipulan datos de modo doloso.

En cuanto a la relación con la administración de justicia, el texto no erige un capítulo exclusivo, pese a que la IA en sede judicial representa un escenario de altísimo riesgo para el debido proceso. El articulado se limita a principios generales de transparencia y no discriminación, sin ofrecer un apartado especial que regule la implementación de sistemas de IA en procesos jurisdiccionales, con pautas de explicabilidad reforzada y salvaguardas ante la automatización de decisiones. Si bien esta omisión no anula la aplicabilidad de los principios éticos al sector judicial, presumiblemente se requerirá una directriz o adaptación posterior que armonice estas reglas con la potestad jurisdiccional y la independencia del juez.

Hacia el cierre, el proyecto subraya la importancia de la ciberseguridad (artículo 38) y la no discriminación en el ámbito laboral (artículo 40). Resulta especialmente encomiable la previsión que el reconocimiento facial y otros algoritmos no supongan invasiones arbitrarias a la intimidad o generen efectos discriminatorios. Esta preocupación se alinea con las directrices de la Unión Europea, aunque el texto, nuevamente, no desciende en especificar mecanismos concretos de evaluación *ex ante* de las herramientas biométricas.

En conclusión, la propuesta contenida en el expediente legislativo 23.919 articula un andamiaje normativo ambicioso que, si bien enuncia principios actualizados y apuesta por la interoperabilidad y la colaboración multiactor, todavía exhibe lagunas en la clasificación de riesgos, en la precisión de las sanciones y en la determinación de procedimientos de control

sustantivos. Su mayor virtud radica en la conjunción de un discurso promocional —que impulsa la investigación, la innovación y las alianzas— y un talante garantista que no rehúye incorporar dimensiones éticas, de rendición de cuentas y de respeto por los derechos fundamentales y el medio ambiente. Para robustecer la eficacia real del texto en un entorno tan complejo como la inteligencia artificial, resultaría aconsejable delimitar, con mayor nitidez, las escalas de riesgo, consolidar un régimen sancionador graduado y desarrollar una norma complementaria que dote de rigor técnico a la evaluación de impacto y a la explicación algorítmica. Con tales modificaciones, la “Ley para la Promoción Responsable de la Inteligencia Artificial en Costa Rica” poseería los atributos necesarios para insertarse, sin desmedro, en la tradición costarricense de equilibrio entre la innovación tecnológica y la primacía incuestionable de la persona humana.

- **Proyecto de Ley N.º 24.484: Ley para la Implementación de Sistemas de Inteligencia Artificial (IA)**

Desde su exposición de motivos, este texto legislativo ofrece un panorama inicial que refiere no solo al desarrollo histórico de la inteligencia artificial —cuyos antecedentes remontan a la década de los cincuenta del siglo XX—, sino, también, a las preocupaciones y oportunidades que la IA genera en la actualidad. De esta forma, la exposición preliminar del proyecto no se limita a un recuento científico: traza las implicaciones sociales, económicas y jurídicas de la IA, delineando una propuesta de regulación que pretende asegurar que las tecnologías automatizadas obedezcan a principios de humanidad, transparencia y rendición de cuentas.

El proyecto exhibe la confluencia de cuatro grandes ejes temáticos. En primer lugar, plantea la pertinencia de acudir a precedentes comparados, evidenciados en normas de reciente promulgación, especialmente en la Unión Europea, en diversos estados de los Estados Unidos y en países latinoamericanos como Perú. Este diálogo con la experiencia legislativa de otras jurisdicciones conduce a la incorporación de conceptos como la responsabilidad de los proveedores de IA, la categorización del riesgo en la implementación de sistemas automatizados y la necesidad de establecer salvaguardas éticas en ámbitos sensibles (educación, salud, fronteras, entre otros). En segundo término, el texto incorpora la exigencia de salvaguardar la propiedad intelectual y los derechos de imagen ante el surgimiento de herramientas de IA generativa capaces de crear contenido nuevo —textos, músicas, audiovisuales— a partir de la ingesta de datos o de

materiales protegidos por derechos de autor. Como tercer eje, consagra la aspiración de elevar los estándares de transparencia y de supervisión humana, al establecer procedimientos de evaluación *ex ante* en ciertos sectores y la obligatoriedad de etiquetar y advertir al usuario acerca de la presencia de sistemas de IA. Por último, hace énfasis en la protección de derechos fundamentales, con particular atención a los datos personales, la no discriminación, la integridad de la información, así como la prohibición de ciertas aplicaciones de la IA que podrían vulnerar derechos o lesionar la autonomía humana.

La propuesta legal se organiza en un cuerpo articulado que arranca con el Capítulo I, centrado en el ámbito de aplicación y en la competencia territorial y material de la ley. En este apartado, el legislador aspira a abarcar, tanto a las personas físicas, como jurídicas, que ofrezcan sistemas de IA desde fuera del país, siempre que estos estén al alcance de usuarios en Costa Rica. Tal extraterritorialidad funcional es congruente con la naturaleza transnacional de los servicios digitales y se fundamenta en la lógica de la “afección”: si un producto o servicio de IA despliega efectos en territorio costarricense, su operador cae bajo la jurisdicción nacional. De modo concomitante, se impone responsabilidad solidaria a todo aquel que suministre los datos o las bases de datos empleados para entrenar los sistemas de IA en cuestión, con la finalidad de impedir la elusión de la norma.

El texto avanza luego hacia la definición de principios rectores y derechos de los usuarios y grupos afectados por la IA (Capítulos I y II). Allí se describen nociones que en la dogmática actual sobre inteligencia artificial constituyen elementos vertebradores: la centralidad de la persona humana, la supervisión humana adecuada, la protección de la privacidad, la seguridad, la transparencia y la no discriminación. Se enfatiza, asimismo, la necesidad de garantizar los derechos de corrección en caso de resultados discriminatorios o sesgados y la obligación de rendir cuentas por parte de los desarrolladores y operadores de IA. Este bloque de principios y derechos se enmarca en una concepción robusta de la dignidad humana, en la que el diseño, desarrollo y despliegue de tecnologías automatizadas debe estar alineado con el orden público y la moral social.

El Capítulo III, dedicado a las definiciones, tiene especial relevancia por la claridad conceptual que ofrece a la hora de delimitar términos como “inteligencia artificial”, “sistema de inteligencia artificial”, “IA generativa” y “IA predictiva”. En la dinámica jurídica, contar con

definiciones positivas evita que la ambigüedad frustre los objetivos de la ley y facilita la comprensión común de los ámbitos a regular. Concretamente, la ley estipula la posibilidad de que haya IA de distintos tipos: la generativa, la cognitiva y la predictiva, con el fin de abarcar un repertorio amplio de tecnologías que, en el día a día, no se restringen a un único modelo de interacción.

A continuación, en el Capítulo IV, la norma aborda de manera sistemática aquellas áreas que denomina “zonas de impacto primario”, sujetas a evaluación y autorización por parte del Estado. El legislador costarricense define un listado de actividades consideradas de alto riesgo, entre las que se incluyen la manipulación de datos personales, la salud, la educación, la infraestructura crítica, el sistema judicial, las campañas electorales y los servicios dirigidos a menores de edad. Este enfoque sectorial, que bebe en parte de los estándares y las propuestas reguladoras de la Unión Europea, se traduce en la exigencia de un control previo —por intermedio de la autoridad administrativa competente— antes de que se desplieguen soluciones de IA que puedan acarrear daños irreparables a individuos o colectivos vulnerables. En otras palabras, el proyecto asume la importancia de prevenir prácticas peligrosas o abusivas mediante un proceso de evaluación que, en teoría, debería equilibrar la protección de los derechos y la promoción de la innovación.

Mención aparte merece la regulación sobre identificación biométrica en tiempo real que, con base en lo propuesto, únicamente podría llevarse a cabo por las autoridades nacionales competentes y exclusivamente cuando se investiguen delitos graves con orden judicial. Esta disposición, inspirada en las restricciones internacionales a la vigilancia masiva, pone de relieve la preocupación por la posible invasión a la intimidad ciudadana y por el escalamiento del control estatal mediante sistemas de reconocimiento facial. Asimismo, en el texto se incluyen prohibiciones claras sobre la generación de contenidos falsos (*deepfakes*) sin la autorización del titular de la voz o la imagen y la explotación de obras bajo derecho de autor sin los debidos permisos. Estas interdicciones responden al fenómeno creciente de la desinformación y al peligro de suplantación de la identidad a través de herramientas de IA generativa, salvaguardando la autonomía personal y la propiedad intelectual.

Una de las secciones más significativas, tanto para la realidad costarricense, como para cualquier estudio centrado en la IA y la administración de justicia, es la que atañe al alcance de la prohibición de “elaborar sentencias o decisiones” del Poder Judicial mediante IA. El artículo 11 dispone expresamente que ningún sistema de inteligencia artificial puede reemplazar o siquiera asumir la función de dictar resoluciones judiciales, reglas de derecho o proyectos normativos en manos de los otros poderes del Estado. Esta restricción, si bien se inspira en el propósito legítimo de proteger la independencia judicial y el núcleo irreemplazable del criterio humano, también lleva a un debate profundo: una interpretación demasiado rígida de la disposición podría inhibir la adopción de IA como herramienta de apoyo analítico para redactar borradores, sistematizar jurisprudencia o evaluar patrones de casuísticas repetitivas. En el ámbito de los procesos jurisdiccionales, la vanguardia técnica sugiere que la IA, empleada bajo supervisión de los jueces, podría agilizar la gestión procesal y contribuir a la transparencia y uniformidad de criterios. Sin embargo, la forma en que el texto articula la prohibición no hace diferenciaciones matizadas entre “reemplazo total de la función decisoria” y “apoyo o asistencia tecnológica”. De adoptarse una interpretación extensiva, se correría el riesgo de frenar todo tipo de innovación que facilite la labor judicial, lo cual constituiría una oportunidad perdida para mejorar la eficiencia y reducir la mora judicial.

Desde esa perspectiva, la regulación costarricense lograría armonizar la no sustitución del juez humano con la adopción de mecanismos de IA si, en el posterior desarrollo reglamentario, se establecen distinciones claras que permitan el uso de IA solo como soporte orientativo, siempre supeditado al criterio final del juzgador. Esa apertura no aparece explícitamente en el proyecto y, en consecuencia, su definición quedará presumiblemente en manos del Reglamento Ejecutivo y de la práctica concreta del Poder Judicial, con el acompañamiento de la autoridad designada.

En el Capítulo VI, la norma se detiene en las obligaciones comunes que deben observar los proveedores y operadores de IA de uso general. Se establece la obligación de contar con políticas internas de buenas prácticas, protocolos de confidencialidad y protección de datos, así como un requisito de rendir informes claros que detallen las fuentes utilizadas para entrenar los sistemas. Dicho requisito, que abarca, tanto las obras protegidas por derechos de autor, como datos de carácter personal, refuerza la idea de un ecosistema de IA más transparente y auditabile. Este tipo de obligación se interpreta como una manera de equilibrar el campo de juego entre titulares de

derechos y desarrolladores de tecnología, pues dificulta la apropiación indiscriminada de contenido ajeno y reduce la opacidad algorítmica.

El Capítulo VII, por su parte, incide en la implementación y la gobernanza. El texto ubica en la Dirección de Investigación, Desarrollo e Innovación del MICITT el eje rector de las políticas estatales en la materia. Esta dependencia gubernamental se encargaría de todos los procedimientos de evaluación *ex ante* para las actividades de alto riesgo, pero también de fungir como asesor de las diversas dependencias del Estado en lo atinente al diseño de reglamentos y al esclarecimiento de controversias. El diseño institucional, tal como se presenta, aspira a dotar a esta Dirección de recursos suficientes (sea a través de transferencias directas del presupuesto nacional o por medio de donaciones) para que pueda sostener iniciativas de formación, auditoría y vigilancia sobre el uso de la IA. Su efectividad dependerá, naturalmente, de la precisión y el detalle con los que se aborden estos mandatos en el reglamento de la ley, así como de la asignación real de personal técnico especializado y medios tecnológicos adecuados para ejercer la labor de fiscalización.

Las disposiciones transitorias, por último, establecen un plazo de seis meses para que el Poder Ejecutivo emita el reglamento y confieren a la administración la potestad de delinejar los aspectos procedimentales inherentes a la autorización y evaluación de los sistemas de IA. Con base en la experiencia comparada, se anticipa que tales reglamentos resultarán cruciales para concretar los estándares y requisitos con los que deberán cumplir los desarrolladores, las empresas que oferten servicios de IA y las instituciones públicas ansiosas de implementar estas tecnologías.

Considerados en su conjunto, los aspectos más fuertes del proyecto residen en la incorporación de principios que sintonizan con la vanguardia internacional: la protección de derechos fundamentales, la transparencia, la diferenciación de niveles de riesgo y la imposición de un control estatal previo en casos de alto impacto social. Asimismo, su énfasis en la rendición de cuentas y la trazabilidad de los datos empleados para entrenar los modelos de IA suponen un avance en materia de información al usuario y a los titulares de derechos. Por otro lado, la extensa prohibición que el texto impone al uso de la IA en la emisión de sentencias judiciales —así como su efecto posible en inhibir la adopción de sistemas automatizados de apoyo— representa quizás el punto más polémico y susceptible de un desarrollo más matizado. En el plano práctico, la comunidad jurídica y tecnológica tendrá que esforzarse en conciliar dos objetivos: impedir la

sustitución de la conciencia humana en la impartición de justicia y, al mismo tiempo, fomentar la modernización y agilidad que la IA brinda a la administración judicial.

En términos de interacción con la temática de la implementación de IA en procesos jurisdiccionales, el proyecto de ley refleja una tensión comprensible: por una parte, se reconoce el alto riesgo que representa delegar decisiones judiciales a entidades informáticas que, por su naturaleza, carecen de la empatía, la prudencia y la responsabilidad social indispensables en la adjudicación de derechos; por la otra, se abre la puerta a la necesidad de integrar herramientas digitales avanzadas que permitan reducir la congestión de los tribunales, sistematizar jurisprudencia y mejorar la eficiencia en la tramitación de causas. El artículo 6, en la medida en que sitúa al “sistema judicial” como actividad sujeta a evaluación previa, habilita la posibilidad de recurrir a IA como instrumento auxiliar si —y solo si— cumple con los parámetros que fijará el MICITT y con la prohibición contenida en el artículo 11. Esa combinación de normas deja clara la voluntad de no erradicar la tecnología, sino de subordinarla a un control humano pleno.

Concluyendo, esta propuesta legislativa costarricense constituye un intento relevante por colocar al país en la senda de la innovación responsable, con el doble propósito de capitalizar los beneficios económicos y sociales de la IA y de proteger, al mismo tiempo, los derechos de la persona. En atención a la temática jurisdiccional, se trata de un marco normativo que, si bien protege la primacía de la función jurisdiccional humana, corre el riesgo de ser interpretado de manera tan estricta que restrinja la colaboración tecnológica. Se espera, por tanto, que el reglamento y las prácticas interpretativas futuras refinen los contornos de estos artículos, de modo que se preserve la garantía de una justicia impartida por personas, al tiempo que la IA pueda servir como herramienta eficaz en el análisis de datos, la gestión de expedientes, la prevención de errores y la administración de la carga judicial.

El proyecto de ley, así concebido, exhibe fortalezas notables en materia de coherencia con normas extranjeras, equilibrio entre innovación y derechos fundamentales y una minuciosa regulación de las áreas más sensibles. No obstante, para convertirse en un instrumento verdaderamente dinámico y efectivo, requerirá de un reglamento que perfeccione las disposiciones sobre el uso de la IA en la judicatura, clarifique las exigencias técnicas y promueva la instauración de canales de colaboración institucional.

En síntesis, Costa Rica cuenta ya con una **estrategia nacional vigente** que guía la IA en el sector público (ENIA 2024-2027) y avanza en la discusión de un **marco legal específico** que podría consolidarse en el corto o mediano plazo. Estas políticas generales –promovidas por el Poder Ejecutivo vía MICITT y complementadas por iniciativas legislativas– tendrán un efecto directo sobre la administración de justicia. Al establecer estándares y lineamientos transversales para el uso de IA (en materia de transparencia de algoritmos, evaluación de riesgos, rendición de cuentas, etc.), obligarán a los actores judiciales a adecuarse a ellos. A la vez, al proveer recursos estratégicos (centros de excelencia, capacitaciones, inversión en infraestructura), facilitarán que el Poder Judicial cuente con **mejores herramientas y conocimiento** para aprovechar la IA en sus funciones misionales.

3.2.- Vacíos y Necesidades de Reforma

El avance creciente de la inteligencia artificial (IA) en Costa Rica y su incipiente adopción en ámbitos judiciales exigen un análisis meticuloso sobre la adecuación de nuestro ordenamiento jurídico ante las transformaciones que conlleva la “revolución algorítmica”. Tal exigencia dimana no solo de una perspectiva de eficacia institucional (agilidad, gestión de mora judicial, etc.) sino, sobre todo, de la ineludible misión de preservar los principios fundamentales que sostienen el Estado de Derecho. A la luz de lo expuesto en apartados precedentes, salta a la vista que la introducción de sistemas automatizados de toma de decisiones en la judicatura pone en jaque elementos nucleares como la independencia judicial, la tutela judicial efectiva, la protección de datos personales y la no discriminación.

En este epígrafe se examinan, en profundidad, los vacíos, las carencias y deficiencias del ordenamiento costarricense en materia de IA, así como las necesidades de reforma para allanar la adopción de sistemas de IA en la administración de justicia de manera compatible con los valores constitucionales. Con tal fin, se aborda el tópico desde tres ángulos específicos: (i) brechas normativas, (ii) aspectos procesales a modernizar y (iii) requisitos institucionales y organizativos

3.2.1. Brechas Normativas Identificadas

La introducción progresiva de sistemas de inteligencia artificial (IA) en el ámbito de la administración de justicia costarricense se topa con un entramado jurídico que, si bien en

determinados puntos, ofrece principios valiosos para la defensa de derechos fundamentales (como las garantías procesales, la protección de datos personales o la prohibición de discriminación), todavía exhibe una ausencia de disposiciones legales expresas y concretas para regular estas nuevas tecnologías. Dicha carencia se percibe con mayor nitidez cuando se examinan las múltiples aristas de la IA aplicada a la función jurisdiccional: la necesidad de supervisión humana, la posible “automatización” de ciertas decisiones, la protección de la independencia judicial, la transparencia de los algoritmos, la responsabilidad civil o disciplinaria en caso de fallas técnicas, la prevención de sesgos discriminatorios, y un largo etcétera. Al no existir una norma sectorial que articule de forma holística los requisitos, obligaciones y salvaguardas respecto a la IA en la judicatura, la incursión de herramientas algorítmicas se conduce, en la práctica, con base en interpretaciones extensivas de principios constitucionales y legales generales, pero sin un marco reglado que brinde seguridad jurídica a los operadores.

Desde una mirada amplia, podría argumentarse que no solo hace falta una regulación sectorial puntual sobre “IA judicial”, sino, también, la expedición de una norma global o transversal sobre la IA —parecida a las que algunos países han promovido recientemente— que incluso de manera tangencial o indirecta, desde la óptica general del “riesgo” (al estilo del AI Act europeo), cubra las especificidades del entorno judicial. Tal abordaje permitiría, al menos, señalar con nitidez la categoría de “alto riesgo” que supondrían los usos de IA en la función de juzgar y, con ello, imponer requisitos de evaluación *ex ante*, transparencia reforzada, explicabilidad y supervisión humana efectiva. De este modo, los eventuales proyectos de ley que se discutan en la Asamblea Legislativa podrían contemplar no solo la dimensión comercial o productiva de la IA, sino, también, el prisma del rol jurisdiccional del Estado y la tutela de derechos fundamentales frente a la lógica automatizada.

En lo que respecta específicamente a los vacíos normativos, conviene subrayar, en primer lugar, la total inexistencia de un cuerpo normativo sectorial que regule a profundidad la implementación de la IA en los tribunales costarricenses. Aunque el texto de la Constitución y la jurisprudencia constitucional aportan principios tuitivos (no discriminación, independencia judicial, debido proceso, integridad y confidencialidad de la información, etc.), se requiere algo más concreto que indique cómo y bajo qué condiciones la judicatura puede adoptar herramientas de inteligencia artificial, estableciendo, por ejemplo, mecanismos de aprobación previa, auditoría y responsabilidad. La normativa dispersa —como la Ley de Certificados, Firmas Digitales y

Documentos Electrónicos o la Ley de Protección de la Persona frente al Tratamiento de sus Datos Personales— atiende algunos aspectos relevantes (valididad jurídica de lo digital, protección de la intimidad, etc.), pero no se adentra en las aristas específicas de la toma de decisiones automatizadas en sede judicial.

De modo que, al carecer de una ley o reglamento que regule de manera directa la IA en la administración de justicia, surgen dudas sobre la legitimidad, el alcance y el control de tales sistemas en cada etapa procesal. Por ejemplo, si un juzgado de cobro pretende usar un algoritmo de clasificación para despachar con mayor celeridad ciertos expedientes, no existe hoy un lineamiento preciso que indique qué tipo de pruebas de fiabilidad debe superar la herramienta, cómo se certifican sus resultados, cómo se garantiza la ausencia de sesgos, cuál es el derecho de la parte afectada a objetar la clasificación y pedir su revisión por un humano, etc. Todo se sustenta en interpretaciones *ad hoc*, que no siempre alcanzan un nivel de seguridad jurídica adecuado y podrían, en un futuro, enredarse en discusiones constitucionales.

En segundo término, se percibe un vacío respecto de un régimen legal de clasificación de los usos de IA en la judicatura según su nivel de riesgo. De la experiencia europea emergen claras lecciones: la IA aplicada a funciones de policía, justicia penal o resolución de conflictos se considera de “alto riesgo” y, por ello, su adopción queda sujeta a salvaguardas más estrictas (evaluación previa, registro, supervisión humana ineludible, test reforzado de no discriminación). En Costa Rica, pese a que el Poder Judicial ha empezado a explorar la IA en la clasificación de escritos o la anonimización de sentencias, ningún texto legal ni directriz interna establece qué usos podrían considerarse “bajos” o “altos” en términos de afectación potencial de derechos. Al no existir un estándar escalonado, se corre el riesgo de: (a) tratar por igual aplicaciones de IA triviales (por ejemplo, la selección de jurisprudencia) y otras que inciden profundamente en la decisión final (una recomendación sobre la duración de la pena, el otorgamiento de un beneficio penal o la priorización de expedientes que involucran bienes de gran valor); (b) desconocer que ciertos escenarios exigen supervisión reforzada, protocolos de calidad de datos y mecanismos de explicabilidad, mientras en otros bastarían requisitos más livianos. En consecuencia, la falta de un enfoque normativo basado en el riesgo impide la adopción de la IA más compleja con garantías adecuadas o, por el contrario, conduce a su adopción sin los candados exigibles en un Estado de Derecho.

Otro vacío importante yace en la carencia de disposiciones procesales específicas que articulen el “derecho a cuestionar” los *outputs* algorítmicos y la obligación de la judicatura de motivar cómo integra tales sugerencias. El debido proceso exige la posibilidad de controversia, contradicción, producción de prueba e impugnación razonada de cualquier elemento que influya en la resolución. Si un juzgado penal, por ejemplo, se apoya en un software predictivo que calcula “riesgo de reincidencia” para dictar medidas cautelares, resulta vital que el imputado o su defensor sepan que ese algoritmo se está usando, puedan recabar información sobre sus parámetros, exigir la verificación de la fiabilidad y eventualmente rebatirlo. Sin embargo, las normas procesales vigentes no contemplan un cauce ad hoc para ese escrutinio. Se puede, en teoría, interponer recursos de nulidad, pero todo ello se vuelve excesivamente incierto y podría redundar en litigios prolongados o dejar, en la práctica, indefenso al ciudadano frente a la “opacidad” algorítmica. Lo mismo acontece con la motivación de la sentencia: hoy la ley obliga al juez a exponer sucintamente los hechos, la prueba y el fundamento jurídico, pero nada dice acerca de cómo motivar el uso de un informe estadístico o una recomendación generada por IA. El riesgo es que la decisión final se “encubra” en un dictamen tecnológico difícil de rebatir, difuminando la responsabilidad del juzgador y lesionando la autonomía interpretativa que exige la independencia judicial.

En cuarto lugar, se echa en falta un régimen claro de responsabilidad civil, administrativa e incluso penal cuando se produzcan errores, violaciones de la privacidad o daños a los derechos de las partes ocasionados por una aplicación inadecuada de la IA. Desde el ordenamiento costarricense, la responsabilidad por el “mal funcionamiento de los servicios públicos” recae sobre el Estado, que podría repetir contra el funcionario causante del perjuicio en caso de dolo o culpa grave. Pero en un entorno de IA, ¿cómo se determina la culpa de un juez que confía en un algoritmo deficiente, o la culpa de la empresa proveedora del software? ¿Cómo incide la opacidad del modelo, especialmente si está protegido por secretos comerciales? Si la persona afectada no puede acceder al código fuente ni entender su lógica, resultará complejo asignar responsabilidad. Más aún, ¿podría la judicatura alegar que la falla es propia del desarrollo algorítmico y responsabilizar al proveedor? Estos interrogantes reclaman reformas legales que tipifiquen y regulen la distribución de responsabilidades de manera acorde con la realidad algorítmica. De lo contrario, cabe el peligro de “zonas grises” que dificultan a las víctimas obtener indemnización o remedio.

Por otro lado, en materia disciplinaria, la adopción de IA sin la diligencia debida o condescendiendo a un “sesgo de automatización” que afecte la imparcialidad podría merecer

sanciones para el juez o para la institución, pero la normativa orgánica e interna del Poder Judicial no contemplan tales supuestos específicos. Aunado con ello, no se ha desarrollado un marco conceptual que oriente a los jueces sobre cuál es su margen de facultad discrecional para aceptar o rechazar la sugerencia de un sistema de IA, ni que establezca los límites que no pueden ser rebasados so pena de incurrir en responsabilidad. El resultado es un limbo en el que la IA podría introducir asimetrías sin un verdadero control, o paralizar la innovación por temor a consecuencias disciplinarias inciertas.

Otro vacío, correlacionado con los anteriores, atañe a la “explicabilidad” y a la “transparencia” algorítmicas. El principio de publicidad y transparencia de las actuaciones judiciales debe conciliarse con la protección de datos personales y la propiedad intelectual de los proveedores del software. La existencia de sistemas en modalidad de “caja negra”, en la que ni siquiera el Poder Judicial tiene acceso pleno a la lógica de inferencia del modelo, deviene conflictiva con la obligación de motivar la sentencia y con el derecho del ciudadano a impugnar una decisión no fundamentada en razones comprensibles. Sin una norma sectorial o general que exija a los proveedores de IA “abrir” sus modelos o proporcionar mecanismos de explicabilidad, la dependencia tecnológica puede dejar a la judicatura maniatada, sin control real sobre la herramienta. Todo ello es agravado por la escasa experiencia en auditorías algorítmicas y la falta de lineamientos estatales que definan qué nivel de transparencia es necesario en estos convenios con empresas de tecnología.

Cabe señalar, igualmente, la falta de una regulación clara que prohíba o establezca restricciones a usos de IA considerados excesivamente invasivos. Los analistas especializados han alertado sobre la posibilidad de que, en el futuro, sistemas de reconocimiento biométrico o de monitoreo conductual se introduzcan en entornos judiciales, por ejemplo, para “detectar mentiras” en sala de audiencias o vigilar reacciones de la parte acusada³²⁰. Mientras que en otras latitudes se discute la eventual prohibición de tales aplicaciones por su alto potencial de invasión a la dignidad y la privacidad, en Costa Rica no hay disposición que delimita si ese tipo de IA es admisible o no en el proceso penal o civil. Ello abre la puerta a que el día de mañana, por la vía contractual, se adquiera alguna tecnología intrusiva sin un debate previo ni una ley que la prohíba o restrinja.

³²⁰ Kaitlin Jackson, “Challenging Facial Recognition Software in Criminal Court,” *The Champion* 44, n.º 1 (enero/febrero de 2020): 14. Recuperado de: https://www.nacdl.org/getattachment/548c697c-fd8e-4b8d-b4c3-2540336fad94/challenging-facial-recognition-software-in-criminal-court_july-2019.pdf

Por último, el carácter difuso de la iniciativa legislativa en materia de IA en Costa Rica — varios proyectos de ley en la corriente parlamentaria, ninguno todavía aprobado — deja en la incertidumbre la instauración de un eventual organismo o autoridad que supervise a nivel nacional el desarrollo y uso de la IA, incluidas las aplicaciones judiciales. Frente a la inminente aprobación de una ley marco en el ámbito general de la IA (con la posible creación de la Comisión Interinstitucional o un ente regulador más robusto), podría verse de modo tangencial la cuestión judicial. Sin embargo, la especificidad del campo jurisdiccional sugiere la pertinencia de incluir, en esa misma ley global, una mención expresa a la IA en la justicia como uso de alto riesgo, sometido a una serie de garantías y validaciones reforzadas. Ello permitiría que, incluso en ausencia de una ley sectorial judicial (o mientras esta no llegue), la ley general prevea una cláusula que regule, al menos, la clasificación de riesgo y los principios de supervisión e independencia del juez. En caso contrario, la IA en la judicatura volverá a quedar relegada, confiada a interpretaciones de un articulado general que no aborda la singularidad de la función jurisdiccional.

Para recapitular, las brechas normativas pueden esquematizarse del siguiente modo:

- (i) Falta de un instrumento legal sectorial sobre IA judicial, con previsiones específicas de supervisión, responsabilidad y salvaguardas;
- (ii) Ausencia de un régimen de clasificación de riesgo, que reconozca la IA en la judicatura como un ámbito de “alto riesgo” y, por ende, la someta a exigencias amplificadas;
- (iii) Ninguna norma procesal específica que habilite la impugnación formal de outputs algorítmicos y exija al juez motivar el grado de influencia de la IA en la sentencia;
- (iv) Vacío en la definición de un régimen de responsabilidad civil y disciplinaria adaptado a la compleja dinámica de la colaboración hombre-máquina, así como en la determinación del alcance de la culpa del proveedor, del Estado o del juzgador;
- (v) Ausencia de pautas sobre transparencia y explicabilidad, lo cual conlleva un riesgo de “caja negra” que atente contra la publicidad procesal;
- (vi) Falta de prohibiciones o restricciones expresas para aplicaciones excesivamente invasivas;
- (vii) Inexistencia, por ahora, de una ley general sobre IA que, al estilo del AI Act europeo, cubra de manera omnicomprensiva (o al menos tangencial) los usos judiciales, contemplando la dimensión del riesgo y la exigencia de control humano.

Todo este panorama, además, impide la acción decidida de quienes dentro del Poder Judicial impulsan mejoras tecnológicas. Aun las voluntades más progresistas se ven frenadas por temores fundados: la incertidumbre legal puede desembocar en cuestionamientos de constitucionalidad, reclamos de arbitrariedad o vulneración de la independencia del juez. Por tanto, la carencia de un soporte legal claro, en vez de alentar la innovación responsable, erosiona la confianza en el uso de herramientas inteligentes.

La experiencia europea, con su AI Act —aunque aún en evolución—, enseña la necesidad de un texto legal en el que, partiendo de la tipología de riesgos, se asignen obligaciones de gobernanza de datos, documentación técnica, supervisión humana, registro de logs, análisis de sesgos y, en el caso de IA judicial, un plus de garantías por el impacto en la libertad y los derechos fundamentales de los litigantes. Algo análogo, guardando las proporciones y circunstancias costarricenses, podría ser la vía idónea: una norma global que, sin perjuicio de desarrollar luego un reglamento judicial sectorial, introduzca las nociones de “alto riesgo”, “sesgo algorítmico”, “explicabilidad” y “responsabilidad solidaria” del proveedor en caso de sistemas que incidan en actos jurisdiccionales. Ello resolvería, al menos parcialmente, la laguna regulatoria, de modo que la judicatura no quedara al borde del vacío legal.

En definitiva, la línea roja reside en que cualquier introducción de IA en la administración de justicia no puede cercenar la autonomía del juez ni los derechos procesales de las partes. Pero para concretar dichas cautelas y evitar la paralización, el ordenamiento costarricense debe colmar sus vacíos normativos, sea mediante la promulgación de una ley sectorial sobre IA judicial o, en su defecto, a través de una ley global de IA que consagre principios y preceptos obligatorios para el uso en la judicatura. En ausencia de tal marco, la implementación a gran escala de herramientas algorítmicas quedará sujeta a la buena voluntad y la autorregulación casuística, sin la firmeza jurídica que brinda la ley, situación que puede derivar en inseguridad jurídica, tensiones constitucionales e, incluso, afectaciones concretas a los derechos fundamentales de la población.

No debe soslayarse, por último, que la construcción de un marco normativo que ampare la introducción de la IA en la justicia conlleva equilibrar intereses diversos: el anhelo de modernizar y agilizar la gestión de los tribunales (reduciendo la mora judicial), la salvaguarda de los derechos de las partes procesales (debido proceso, contradicción, igualdad), la protección de la independencia judicial (no sujeción a presiones externas ni a lógicas “automatizadas” de resolución) y la promoción de la transparencia. Tales intereses no son contradictorios si se regulan

con prudencia. El error sería pensar que la IA solo puede introducirse a costa de sacrificar derechos, o que la preservación de principios judiciales básicos implica renunciar a la tecnología. Es perfectamente factible —y deseable— un diseño legislativo que habilite la IA como instrumento auxiliar, pero amarre su uso a exigencias de supervisión humana, explicabilidad, control de sesgos y un régimen de responsabilidad claro.

Desde luego, cualquier regulación debe concebirse como un texto “dinámico”, con revisiones periódicas, dada la velocidad con que evoluciona la inteligencia artificial. La experiencia en otras latitudes muestra que la ley “no puede quedar rezagada”: un estatuto excesivamente rígido podría quedar obsoleto, mientras que uno demasiado abierto podría implicar inseguridad jurídica. Se requerirá, pues, un equilibrio: la ley debe marcar los grandes lineamientos (clasificación de riesgo, obligaciones de transparencia, deber de supervisión humana, responsabilidades) y el Poder Judicial —junto a la autoridad nacional encargada de la IA— podría desarrollar directrices más pormenorizadas, guías de auditoría y protocolos de uso. Así, se llenaría ese vacío que hoy atenaza la adopción de la IA en los tribunales costarricenses.

En suma, la principal brecha normativa es la inexistencia de una regulación explícita y sistemática sobre el uso de IA en la administración de justicia, en conjunción con la ausencia de una ley general de IA que aborde, aunque sea de modo general, las implicaciones de los algoritmos en actividades tan sensibles como la judicial. Esta laguna se plasma en múltiples dimensiones: falta de un régimen de clasificación de riesgos y escalas de supervisión, ausencia de cauces procesales para contradecir outputs algorítmicos, vacío en la determinación de responsabilidades, falta de disposiciones concretas sobre transparencia y explicabilidad, indefinición respecto de posibles usos excesivamente invasivos, y poca claridad sobre la compatibilidad de la IA con la independencia judicial. Todo ello abona a la urgencia de adoptar, en primer lugar, sea una ley sectorial específica o una ley general de IA con un apartado sobre la justicia, para así clarificar los parámetros básicos y evitar que la administración de justicia transite por una zona gris en la que la innovación digital pueda lesionar la base constitucional de nuestro Estado de Derecho.

3.3.- Lecciones del Modelo Europeo

El escenario tecnológico contemporáneo se caracteriza por la emergencia de sistemas de inteligencia artificial (IA) que exhiben capacidades notablemente avanzadas: algoritmos de

aprendizaje profundo, modelos de lenguaje de gran escala, técnicas de minería de datos, entre otras expresiones. Dichas herramientas, con su potencia analítica y la aptitud de procesar volúmenes ingentes de información, han encendido el interés no solo del ámbito productivo o científico, sino, también, de las instituciones encargadas de la función jurisdiccional. La posibilidad de utilizar IA para, por ejemplo, agilizar la gestión de expedientes, sugerir posibles cursos decisarios o uniformar criterios en materias sensibles, ha abierto perspectivas antes inimaginables. Sin embargo, también plantea inquietudes profundas acerca de la salvaguarda de derechos fundamentales y del equilibrio institucional que se desprende de la naturaleza misma del Estado de Derecho.

Como se ha expuesto, el debate se ha vuelto particularmente intenso en la Unión Europea, cuya tradición jurídica, inspirada en la centralidad de la dignidad humana y en la primacía de los derechos fundamentales, ha forjado un entramado normativo que no rehúye la innovación, pero la somete a un proceso de cribado ético y legal minucioso. Este andamiaje puede definirse —en lo esencial— por la existencia de instrumentos como el *Reglamento sobre IA (AI Act)*, así como de un conjunto de directrices y principios recopilados en la *Carta Ética Europea sobre el uso de la IA en los sistemas judiciales* y otros documentos conexos de carácter programático o de *soft law*. El resultado de estos esfuerzos, que distan de ser meramente declarativos, ha contribuido a perfilar un “modelo europeo” para el uso de inteligencia artificial en la justicia, centrado en el principio antropocéntrico, en la evaluación escalonada de riesgos y en el refuerzo de garantías específicas para evitar la lesión de derechos esenciales y la degradación del orden institucional.

Este apartado se propone desarrollar y profundizar en los fundamentos de ese modelo normativo y en los postulados esenciales que lo articulan, de manera que se evidencie cómo la aproximación comunitaria deviene una fuente de inspiración significativa para cualquier otro país que busque regular la introducción de sistemas de IA en la administración de justicia. Se procederá, pues, a un análisis que pone su acento en el contenido cardinal del corpus europeo, en sus presupuestos conceptuales y en las exigencias prácticas de implementación, con la mirada puesta en la posibilidad de trasladar adaptativamente —y no de forma meramente mimética— tales lineamientos a un entorno constitucional distinto. El propósito es, ante todo, resaltar la lógica y los criterios jurídicos que se han erigido en la Unión Europea y que, al juzgar por su ambición y precisión técnica, están llamados a influir en el debate global sobre la materia.

3.3.1. Principios y Garantías Adaptables

La experiencia de la Unión Europea en la integración de la inteligencia artificial en la administración de justicia evidencia que los pilares axiológicos y jurídicos sobre los cuales se asienta el Estado de Derecho no pueden reducirse a meros planteamientos declarativos, sino que exigen un desarrollo normativo y metodológico de honda sofisticación. Desde esta visión, se ha gestado un conjunto de principios y garantías que, aun hundiéndose sus raíces en el acervo constitucional y en los tratados internacionales de derechos humanos, se han visto transformados y especificados para responder a las condiciones técnicas y dilemas éticos propios de la IA.

La referencia a dichos principios no implica una simple remisión a los valores generalísimos de dignidad, libertad o igualdad, pues estos se ven traducidos en exigencias concretas: la sujeción irrestricta del uso de algoritmos a la tutela judicial efectiva, la consagración de la transparencia y la explicabilidad como requisitos técnicos ineludibles, la distinción entre niveles de riesgo para calibrar la intensidad de los controles o la permanente posibilidad de intervención y supervisión humana en toda decisión que pueda afectar derechos fundamentales. Tales directrices, arraigadas en la praxis europea, han dado lugar a un marco adaptativo y dinámico, idóneo para un fenómeno tecnológico que evoluciona sin pausa.

➤ Hacia una Concepción Antropocéntrica de la IA Judicial

La inteligencia artificial, entendida en términos muy generales como la construcción de sistemas informáticos que exhiben capacidades de razonamiento o aprendizaje similares a las del ser humano, ha reconfigurado la concepción de la automatización en prácticamente todos los ámbitos productivos. Lo peculiar de su introducción en la justicia radica en el tipo de función que allí se ejerce: la potestad de resolver conflictos y de impartir resoluciones vinculantes es un pilar esencial del Estado de Derecho, cuyo contenido se impregna de principios como la tutela judicial efectiva, la independencia del juzgador y la garantía de que todo acto decisorio sea conforme con las libertades y los derechos de la persona. Frente a esta realidad, el primer gran pilar del modelo europeo es la idea de que la IA, por más avanzada que sea, no disuelve la centralidad del operador humano en las etapas en que está en juego la valoración estrictamente jurídica o el juicio valorativo que conduzca al pronunciamiento final.

El AI Act y la Carta Ética Europea ponen un énfasis inequívoco en que, en la órbita judicial, los algoritmos han de concebirse como asistentes, catalizadores de eficiencia o instrumentos de apoyo, sin llegar a desplazar el núcleo esencial del juicio humano. De ahí que se utilice con frecuencia la expresión de “antropocentrismo tecnológico”: antes que concebir a la máquina como un sustituto del magistrado, la normativa la ubica como una herramienta programada para servir a la persona, y no al revés.

La importancia de este principio se explica por la íntima relación entre la actividad jurisdiccional y la garantía de los derechos. En el modelo europeo, la dignidad humana aparece siempre como premisa que justifica la imperiosa necesidad de mantener un control de racionalidad y de humanidad en el acto decisorio. No se trata, por tanto, de un freno a la innovación digital, sino de un posicionamiento ético que vertebrá todas las normas e impone que la IA no se convierta en un “oráculo” ante el cual el juez se incline automáticamente, sino en un instrumento sujeto a verificación, cuestionamiento y corrección.

La concretización de este postulado antropocéntrico —y en esto la Unión Europea ha sido enfática— solo adquiere verdadera efectividad si se refrenda con pautas jurídicas concretas en cada una de las fases de la cadena de valor de la IA judicial: diseño, entrenamiento, despliegue, uso en los expedientes y supervisión *a posteriori*. Así, por ejemplo, una de las obligaciones primordiales que se desprende de los documentos europeos es la necesidad de dotar de una interfaz de usuario comprensible al juzgador, de modo que este no se halle forzado a aceptar ciegamente la salida del sistema, sino que pueda interpretarla y, en caso de advertir falencias, apartarse de ella.

Se erige, en consecuencia, la figura de la “supervisión humana efectiva”: no basta con la mera formalidad de que exista un juez; se requiere que dicho juez tenga a su disposición las herramientas cognitivas y técnicas necesarias para comprender cómo la IA ha llegado a su recomendación o pronóstico. Este principio enlaza con el valor más amplio de la “explicabilidad”, tan repetidamente aludido en los lineamientos europeos. Ese afán de asegurar la comprensión humana del funcionamiento algorítmico bebe, por otro lado, de un segundo principio complementario: el de la no sustitución de la potestad jurisdiccional. Ni el AI Act ni la Carta Ética introducen una prohibición absoluta de aplicar IA, sino que impiden suplantar la evaluación prudencial que es inherente al juez. Tal salvaguarda se hace visible, por ejemplo, en la exigencia de proponer mecanismos de rendición de cuentas y de plena trazabilidad en las deliberaciones que inciden en el veredicto. De ahí que sea posible afirmar que la aproximación comunitaria a la IA judicial no se

limita a un tono laudatorio de las virtudes tecnológicas, sino que las integra como parte de un esquema de control y responsabilidad.

➤ **El Enfoque Basado en el Riesgo: Graduación Normativa y Proporcionalidad**

Un segundo principio rector que orienta la estrategia de la Unión Europea es el “enfoque basado en el riesgo”. El legislador comunitario, consciente de la multiplicidad de escenarios en que puede irrumpir la IA y de la dispar trascendencia que cada uno conlleva, delineó un método de clasificación que asigna diferentes categorías según el nivel de amenaza o impacto que esos sistemas puedan ocasionar a los derechos fundamentales o a la seguridad pública.

La gran originalidad de este planteamiento consiste en que no todas las aplicaciones tecnológicas se someten al mismo grado de escrutinio. Aquellas que, por su naturaleza, revisten una incuestionable neutralidad y un riesgo prácticamente nulo —por ejemplo, los sistemas de IA que clasifican de manera asistencial documentos de archivo— se ubican en un nivel de riesgo “mínimo o inexistente”, lo que acarrea tan solo un puñado de obligaciones suaves de transparencia y buenas prácticas. Otros casos, con un potencial moderado de impacto, quedarán sujetos a deberes incrementados, en materia de información al usuario o de supervisión *ex post*. Sin embargo, los usos denominados de “alto riesgo”, capaces de incidir en la decisión judicial o de condicionar de algún modo la libertad de las personas, se subordinan a reglas mucho más estrictas. Y, en ciertos supuestos taxativamente enunciados que sean intrínsecamente incompatibles con la salvaguarda de los derechos de la persona o con la dignidad humana, el propio AI Act prevé la prohibición lisa y llana de su uso.

Aplicado al sector judicial, este método de graduación hace que la IA sea objeto de exigencias cualitativamente superiores cuando se utiliza para tareas que pueden influir en la concepción sustantiva de la sentencia o de la resolución. Así, no se equipará un software de simple gestión logística con un sistema que formule recomendaciones en materia penal o que antice la probabilidad de reincidencia. El AI Act identifica, dentro del anexo correspondiente, la administración de justicia como ámbito de alto riesgo, precisamente por la relevancia que los actos judiciales revisten para la garantía de los derechos fundamentales.

El método de la proporcionalidad tecnológica, más allá de su interés conceptual, se plasma en disposiciones concretas: el desarrollador de sistemas de IA con vocación de incidir en decisiones judiciales debe, antes de la comercialización, someter a auditoría y verificación su modelo, elaborar documentación técnica exhaustiva, prever mecanismos de explicación y permitir registros de actividad que doten de transparencia al proceso algorítmico. Asimismo, con posterioridad a la implementación, las autoridades competentes se hallan autorizadas para exigir informes regulares y, en caso de detectarse fallas serias —por ejemplo, sesgos discriminatorios o índices de error excesivos—, se faculta la adopción de medidas correctivas o incluso la prohibición de uso.

Si se contemplara la posibilidad de introducir un enfoque semejante en nuestro ordenamiento, la conclusión inmediata sería la necesidad de efectuar un mapeo realista de las aplicaciones existentes y potenciales de la IA en la justicia, clasificándolas en función del riesgo. De ello derivarían escalas de obligaciones y salvaguardas adaptadas al perfil de cada herramienta: muy ligeras en caso de sistemas de ayuda documental, moderadas en casos de herramientas de predicción de tiempos procesales y rigurosas cuando la IA pudiera condicionar la dimensión sustantiva de la sentencia. Esta gradación no debe interpretarse como un obstáculo, sino como la posibilidad de alinear la innovación con la prevención de daños a la esfera constitucional.

➤ **La Transparencia y la Explicabilidad como Ejes Transversales**

Uno de los retos más colosales en la adopción de la IA en el ámbito judicial es la necesidad de dotar de transparencia y explicabilidad a los sistemas, especialmente ante el auge de algoritmos de aprendizaje profundo que, por su propia complejidad interna, operan como “cajas negras” poco inteligibles incluso para sus creadores. El legislador europeo, en respuesta a esta dificultad, ha dejado clara la premisa de que no se puede delegar en la máquina la construcción de una decisión con repercusiones jurídicas sin que exista un correlato de trazabilidad y conocimiento público —o, al menos, supervisado— de los parámetros esenciales.

Tal aspiración se plasma en la consagración de estándares como la obligación de documentar la lógica fundamental de la IA, de mantener registros de las etapas más relevantes del proceso de cálculo o de apoyar la interpretabilidad de los resultados a fin de que el usuario perciba

en qué se sustenta la predicción u orientación dada. Esto no significa que haya que publicar irrestrictamente la totalidad del código fuente, pero sí que se facilite una comprensión funcional de las hipótesis, las variables y el peso asignado a cada factor. El AI Act, por consiguiente, exige que los proveedores describan la arquitectura y los métodos de entrenamiento en un formato que las autoridades y el operador judicial puedan analizar.

En sede judicial, la importancia de esta línea de acción radica en que el derecho de defensa y la posibilidad de recurrir una resolución se verían gravemente devaluados si la parte afectada no entendiera los elementos clave del algoritmo que ha influido en su situación jurídica. La tradición europea entiende que la persona tiene derecho a que la decisión que la incide sea explicada de modo suficiente y a que la instancia judicial se exprese con racionalidad. Otorgar un papel significativo a la IA, pero sin proveer la base explicativa, conllevaría un peligro de opacidad inaceptable. Por ello, la regulación comunitaria reitera que las soluciones de alto riesgo operen con un plus de claridad, habilitando la posibilidad de cuestionar y rebatir las conclusiones.

En consecuencia, la divulgación de la lógica esencial de la IA se convierte en una pieza inescindible del debido proceso. No se trata de un mero estándar técnico, sino de una exigencia que, en la experiencia europea, aparece íntimamente ligada al respeto de los principios fundamentales. El ordenamiento prescribe de esta forma la adopción de metodologías de interpretabilidad: que el sistema cuente con bloques de explicación, paneles de control y otros dispositivos capaces de exponer, de manera inteligible, la composición de su dictamen.

➤ **La Supervisión y el Control Humano como Barrera Infranqueable**

El AI Act, en su planteamiento, distingue con nitidez los usos de la IA donde la máquina se limita a asesorar y aquellos donde, hipotéticamente, podría desplazar la última palabra del ser humano. El modelo europeo veta la desplazabilidad total del juez en los procedimientos judiciales, no por una noción romántica de la magistratura, sino porque la independencia y legitimidad del órgano jurisdiccional exige un anclaje personalista que la tecnología no puede suplantar. En otras palabras, la máquina puede ayudar al juez a clasificar escritos, a comparar patrones jurisprudenciales o a sugerir esbozos resolutorios, pero la resolución final ha de provenir de la inteligencia deliberativa de la persona.

Este rasgo esencial se refuerza con la noción de supervisión, entendida como la facultad y el deber del operador humano de intervenir en cualquier momento para —si lo estima oportuno— desoír la recomendación. En la práctica, la supervisión se concreta en la obligación de:

1. Garantizar que el sistema de IA ofrezca una interfaz apta para que el juez vea y entienda los resultados.
2. Permitir corregir o rechazar las sugerencias, con total autonomía decisoria.
3. Conservar un registro de los pasos seguidos y del rol jugado por el componente automatizado en la decisión.

La Carta Ética europea también insiste en que el despliegue de IA en la justicia no se limite a lo meramente normativo, sino que se acompañe de procesos de formación para los funcionarios y operadores, de modo que adquieran competencias tecnológicas que les permitan ejercer un control real. El desconocimiento de la lógica de la IA conduciría *de facto* a la aceptación ciega de sus recomendaciones, malogrando el principio de control humano. El hilo conductor de la regulación busca, pues, invertir la inercia: se pretende otorgar a los magistrados y auxiliares un margen de maniobra que evite la genuflexión frente a lo que el algoritmo sugiere.

Si en la antigüedad la labor de un tribunal se apoyaba únicamente en la lectura de la ley y en el estudio de la doctrina, la era digital conlleva una diversificación de las fuentes de información y la introducción de mecanismos algorítmicos. Sin la suficiente capacitación, no habría un control efectivo, sino solo un formalismo.

➤ **Métodos de Prevención y Reparación de Sesgos**

En la Unión Europea, la integridad y la equidad del proceso judicial son valores tutelados por una constelación de normas. La aplicación de algoritmos de IA que trabajen con datos históricos masivos entraña el peligro de cristalizar estereotipos o traducir a escala computacional las disparidades arraigadas en la realidad social. Se trata de uno de los riesgos más criticados de la automatización en la justicia: la probabilidad de que el juicio humano, capaz de introducir matices y consideraciones particulares, sea sustituido por un juicio estadístico que reproduzca injusticias sistémicas.

La regulación comunitaria enfrenta ese problema mediante múltiples dispositivos. Por un lado, impone, desde la fase de diseño, el análisis de la representatividad de los datos de entrenamiento: se exige a los proveedores de IA que demuestren que los conjuntos de datos no están sesgados en perjuicio de colectivos. Por otro lado, la norma incita a que haya ciclos de retroalimentación en los que se monitoricen las salidas del sistema para comprobar si tienen un impacto desproporcionado en ciertos perfiles. En caso de detectarse indicios de discriminación, se obliga a realizar ajustes de metodología y a revisar los algoritmos. Tal corrección puede conllevar la supresión o reponderación de variables, la inclusión de datos suplementarios que restablezcan un equilibrio o la implementación de técnicas de “desviación controlada” para compensar la infrarrepresentación de grupos vulnerables.

Este mecanismo se ve complementado por disposiciones que reiteran el derecho del individuo a exigir explicaciones y a cuestionar las bases del procesamiento cuando considere que la IA ha actuado con parcialidad. En la práctica, esta prerrogativa está diseñada para que, ante la sospecha de un trato discriminatorio, la parte agraviada pueda forzar una revisión humana y, de ser procedente, la corrección del sistema. Cabe destacar que el legislador europeo, al imponer estas cautelas, no se detiene en la mera prohibición de “discriminar”, sino que articula obligaciones concretas de “gobernanza de datos” y “verificación de impacto”, con el fin de que la prevención del sesgo sea un procedimiento continuo y no una mera fórmula retórica.

➤ Esquema Institucional y Dinámicas de Supervisión

La labor de establecer principios y obligaciones en el texto legal o en la Carta Ética Europea no hubiera sido suficiente sin la existencia de una arquitectura institucional de supervisión y sanción. Así, el AI Act contempla la designación de autoridades nacionales responsables de ejecutar y velar por el cumplimiento de las normas sobre IA. Estas entidades, que ya existen en campos análogos —por ejemplo, los organismos que controlan la protección de datos—, contarán con facultades para inspeccionar los sistemas utilizados en los tribunales, reclamar documentación y, si fuera necesario, imponer multas o proscribir la comercialización de un software que no se ajuste a las exigencias legales.

En la esfera comunitaria, también se destaca la creación de foros de coordinación: la Junta Europea de Inteligencia Artificial, en la que se integran representantes de los Estados miembros y de la Comisión, coordina las interpretaciones y guías. Se pretende así reducir el riesgo de que cada Estado aplique criterios divergentes, lo que socavaría la eficacia de un mercado unificado y la coherencia de la protección a la persona. Tal centralización selectiva no significa el vaciamiento de la soberanía de cada país, sino que obedece al interés de asentar principios comunes y promover la convergencia regulatoria.

➤ **El Dinamismo Normativo ante la Evolución de la IA**

Una característica que se constata en el proceso de elaboración y adopción del AI Act es la conciencia del cambio vertiginoso que atraviesan las tecnologías de inteligencia artificial, particularmente en ámbitos como el aprendizaje profundo y las redes neuronales. El texto europeo (Arts. 4 y 84 del AI Act), lejos de clausurar la posibilidad de ajustes, introduce cláusulas de revisión periódica del régimen, permitiendo a la Comisión y al Parlamento evaluar, en plazos prefijados, la idoneidad de las categorías de riesgo y de las obligaciones impuestas. Así se reconoce que, en un entorno tan volátil, la rigidez legislativa puede envejecer rápidamente.

Este rasgo de flexibilidad refleja la convicción de que la respuesta normativa debe actualizarse al compás del progreso tecnológico: lo que hoy se reputa como alto riesgo —por ejemplo, la IA predictiva en materia penal— podría evolucionar a una fase de mayor confiabilidad si los algoritmos demuestran una tasa de error estadísticamente aceptable y una ausencia sistemática de sesgos. Del mismo modo, nuevas aplicaciones que hoy no se contemplan, surgidas de la experimentación científica, podrían ameritar el traslado a la categoría de usos prohibidos si se advierte un impacto frontal contra la dignidad o la libertad de las personas. De esta manera, la Unión Europea consolida un modelo de *regulación evolutiva*, que no se adscribe a un texto pétreo, sino que mantiene la capacidad de reconfigurar y recalibrar sus mandatos conforme la realidad vaya imponiendo nuevos dilemas.

➤ **Relevancia de los Principios Europeos para el Contexto Costarricense**

La propuesta de erigir un marco normativo en Costa Rica inspirado en los postulados europeos no equivale a una trasposición acrítica de reglas extranjeras. Antes bien, exige un estudio

comparado que pondere, en toda su extensión, la armonía entre los valores rectores del AI Act y la tradición constitucional costarricense, caracterizada por la defensa robusta de los derechos fundamentales, la supremacía constitucional y la autonomía de los poderes públicos.

Sin embargo, hay razones para sostener que los principios del modelo europeo —antropocentrismo, control humano, enfoque de riesgo y no discriminación— se conectan de manera bastante natural con la doctrina y el acervo axiológico del ordenamiento costarricense, el cual ha marcado la primacía de la persona como centro y fin último de la actividad del Estado. El hilo conductor es el mismo: la justicia no puede verse reducida a un procedimiento puramente mecanizado, pues la presencia de matices interpretativos y la dignidad del justiciable exigen una interacción reflexiva.

Por otro lado, la metodología que el AI Act implementa, con las distintas categorías de riesgo y las obligaciones correlativas, podría trasladarse adaptativamente a las realidades procesales e institucionales del Poder Judicial de Costa Rica. La categorización, en vez de pretender uniformar, permitiría conferir un esquema normativo donde cada aplicación sea objetivamente encuadrada y se impongan exigencias más suaves o más intensas, según el nivel de repercusión. Las experiencias de la Unión Europea indican que este procedimiento escalonado ayuda a vencer los recelos y a situar el debate en el plano de la evidencia y no en la mera especulación.

La creación de comités institucionales o la asignación de competencias a una autoridad de supervisión en IA podrían, igualmente, encontrar un correlato en la estructura costarricense, sin que sea estrictamente necesario duplicar las instituciones, sino más bien ampliar las facultades de órganos existentes o constituir un ente específico para la evaluación y auditoría de la IA en el sector público, y muy especialmente en la administración de justicia. Dicho ente, siguiendo la pauta europea, debería ostentar facultades sancionatorias y la pericia técnica suficiente para afrontar un examen real de los algoritmos.

En lo concerniente a la transparencia y a la explicabilidad, la legislación costarricense podría recoger preceptos claros para la “interfaz de usuario” y la “motivación de los actos jurisdiccionales”. De esta forma, se codificaría la obligación de que las soluciones de IA provean

explicaciones inteligibles y de que la decisión del juez que se asiente en un recurso algorítmico explice cómo se integró dicha recomendación en la lógica final. Un desarrollo normativo de tal tenor no solo impulsaría la coherencia interna, sino que reforzaría la confianza de la ciudadanía en la adopción de las nuevas tecnologías.

Finalmente, la parte relativa a los sesgos y la discriminación podría encontrar un sólido fundamento en la cláusula de igualdad presente en la Constitución costarricense, postulándose un deber de “debida diligencia” en la programación y uso de IA para neutralizar cualquier tendencia perjudicial hacia grupos vulnerables. Más allá de la mera prohibición de discriminar, la norma resultante incitaría el análisis continuo de los algoritmos y la instauración de métodos de corrección, de modo muy semejante a lo que Europa denomina “gobernanza de datos y auditabilidad”.

3.3.2. Mecanismos de Control y Supervisión

El análisis minucioso de la normativa sectorial europea en materia de inteligencia artificial y su aplicación en el ámbito judicial, desarrollado en el Capítulo II, arroja un caudal de lecciones y buenas prácticas que pueden servir de valioso referente para el caso costarricense. Destacan, en particular, los robustos mecanismos de supervisión y control que el legislador europeo ha diseñado para garantizar un uso responsable y garantista de los sistemas de IA en un ámbito tan sensible como la administración de justicia.

En primer término, procede recapitular los instrumentos normativos basilares que, en el contexto de la Unión Europea, configuran el marco regulatorio de la IA judicial. El Reglamento sobre Inteligencia Artificial (AI Act) se erige como la piedra angular de este *corpus iuris* emergente. Su vocación omnicomprensiva y su enfoque basado en el riesgo lo convierten en un verdadero "código europeo de la IA", con disposiciones específicas para los sistemas empleados en el ámbito judicial.

La Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales y su Entorno, adoptada en 2018 por la Comisión Europea para la Eficacia de la Justicia (CEPEJ), constituye otro referente ineludible. Si bien carece de fuerza vinculante, este instrumento de *soft*

law enuncia los principios rectores y deontológicos que deben guiar el despliegue de la IA en la justicia, con particular énfasis en la protección de los derechos fundamentales.

El Reglamento General de Protección de Datos (RGPD), plenamente aplicable a los tratamientos algorítmicos de datos judiciales, aporta un estrato adicional de garantías en materia de licitud, lealtad y transparencia, minimización de datos y salvaguardas frente a decisiones automatizadas. Por su parte, las resoluciones del Parlamento Europeo, especialmente la de 6 de octubre de 2021 sobre la inteligencia artificial en el Derecho penal, completan el acervo comunitario con una serie de directrices y recomendaciones de especial relevancia para la IA en la justicia.

Del análisis combinado de estos instrumentos se derivan una serie de mecanismos de supervisión y control que, por su carácter garantista e innovador, merecen especial consideración:

a) Autoridades de control y gobernanza multinivel: tanto el AI Act como el RGPD prevén un sistema de supervisión a cargo de autoridades independientes, tanto a escala europea como nacional. Se establece así una arquitectura institucional multinivel, donde coexisten órganos especializados como el Comité Europeo de Protección de Datos, el futuro Comité Europeo de Inteligencia Artificial, las autoridades nacionales de protección de datos y los eventuales entes nacionales de supervisión de la IA.

Esta diversificación orgánica, lejos de fragmentar el control, busca garantizar un enfoque cooperativo y armonizado, adaptado a las especificidades de cada Estado miembro. En el ámbito judicial, se prevé una estrecha colaboración entre estas autoridades de control horizontales y los órganos de gobierno del Poder Judicial, preservando la independencia jurisdiccional al tiempo que se vela por el cumplimiento de los estándares europeos de protección.

b) Obligaciones reforzadas de transparencia y trazabilidad: el AI Act impone a los proveedores y usuarios de sistemas de IA de alto riesgo, categoría en la que se enmarcan expresamente las aplicaciones en el ámbito judicial, obligaciones ampliadas de transparencia y trazabilidad. Entre ellas destacan el deber de documentación técnica detallada (art. 11), el registro automático de eventos (art. 12), la provisión de instrucciones de uso claras y accesibles (art. 13) y la implementación de medidas de supervisión humana (art. 14).

Estas exigencias de transparencia "por diseño y por defecto" buscan conjurar la opacidad tradicionalmente asociada a los sistemas algorítmicos complejos, sometiendo su desarrollo y uso a un escrutinio permanente. En la justicia, donde la motivación de las resoluciones es presupuesto indeclinable de su legitimidad, esta "transparencia aumentada" deviene crucial para posibilitar un control efectivo de las decisiones algorítmicamente asistidas.

c) Evaluaciones de conformidad *ex ante* y auditorías: para los sistemas de IA considerados de alto riesgo, el AI Act establece un exigente sistema de evaluación *ex ante* de la conformidad (arts. 43 y ss.). Antes de su introducción en el mercado o puesta en servicio, estos sistemas deberán superar un proceso de verificación del cumplimiento de los requisitos establecidos en el Reglamento, ya sea mediante control interno (módulo basado en el sistema de gestión de la calidad) o mediante intervención de un organismo notificado.

En el caso de los sistemas empleados en el ámbito judicial, donde los riesgos para los derechos procesales son particularmente acusados, cabe postular que la evaluación de conformidad a cargo de un tercero independiente debería configurarse como regla general. Ello garantizaría una "auditoría algorítmica" externa y tecnicamente cualificada, capaz de detectar sesgos discriminatorios, disfunciones o vulnerabilidades antes de su aplicación a casos reales.

d) Requisitos estrictos de calidad y gobierno de datos: consciente de que la solvencia de los sistemas de IA depende críticamente de la idoneidad de los datos de entrenamiento y de su correcta gobernanza, el legislador europeo dedica especial atención a esta materia. El AI Act impone a los proveedores de sistemas de alto riesgo la obligación de establecer prácticas de gestión de datos adecuadas (art. 10), que abarcan la evaluación de la calidad, exhaustividad y representatividad de los *datasets*, la identificación y corrección de sesgos potenciales, además de la definición de los formatos y procedimientos de captura y conservación.

Proyectada al escenario judicial, esta exigencia de "higiene del dato" adquiere una relevancia superlativa, habida cuenta del carácter altamente sensible de la información jurisdiccional y de su impacto directo en los derechos de los justiciables. La selección y etiquetado de las resoluciones judiciales utilizadas para entrenar los algoritmos deberá realizarse con la

máxima solvencia metodológica, a fin de conjurar el riesgo de perpetuación tecnológica de prejuicios o pautas decisorias erráticas históricamente arraigadas.

e) Derecho a la revisión humana y a la impugnación de decisiones automatizadas: tanto el AI Act (en sus disposiciones generales sobre supervisión humana) como el RGPD (en su régimen específico para las decisiones individuales automatizadas) convergen en la necesidad de preservar espacios irreductibles de valoración humana y de garantizar vías de recurso efectivas frente a resoluciones íntegramente confiadas a la máquina.

Así, se consagra el derecho de todo interesado a no verse sometido a una decisión basada únicamente en el tratamiento automatizado cuando esta produzca efectos jurídicos que le afecten significativamente (art. 22 RGPD), con la correlativa obligación del responsable de articular mecanismos de revisión e impugnación que permitan rebatirla. En el ámbito judicial, donde toda decisión debe poder ser revisada por un órgano superior, este "derecho al recurso" y a la valoración humana individualizada se erige en garantía crucial de la equidad procesal.

f) Régimen sancionador disuasorio: finalmente, tanto el AI Act como el RGPD, articulan un sistema de sanciones de notable severidad por incumplimiento de las obligaciones en ellos establecidas. Las multas previstas pueden alcanzar, para las infracciones más graves, los 30 millones de euros o el 6 % del volumen de negocio total anual global (Art. 71 del AI Act).

Esta contundencia punitiva, que sitúa las sanciones en AI Act en el umbral superior de las establecidas en el ordenamiento europeo, refleja la firme voluntad del legislador de dotar de efectividad a sus prescripciones mediante una amenaza creíble y financieramente onerosa. Aplicado al sector público, y específicamente al ámbito judicial, este régimen sancionador actuaría como potente incentivo para que las administraciones de justicia implementen con el máximo rigor los estándares de protección requeridos.

La valoración crítica de este caudal de mecanismos de supervisión y control arroja, en su conjunto, un balance netamente positivo en términos de adecuación finalista y proporcionalidad. El modelo europeo aspira a casar la imperiosidad de precauciones reforzadas, justificada por la magnitud de los riesgos en presencia, con la necesaria flexibilidad y adaptabilidad a las circunstancias de cada aplicación concreta. Rehúye, así, tanto la parálisis regulatoria, como el

intervencionismo asfixiante, en pos de un marco habilitante que conjure abusos o desviaciones sin obturar los considerables beneficios que una IA "fiable y controlada" puede reportar a la administración de justicia.

Ciertamente, la traslación de estos mecanismos al ecosistema jurídico costarricense no está exenta de desafíos. La creación de autoridades públicas sectoriales con la cualificación técnica y los recursos bastantes para acometer una supervisión efectiva requerirá un esfuerzo presupuestario y organizativo considerable. La "alfabetización algorítmica" de los operadores judiciales, llamados a officiar de barrera de contención frente a automatismos indeseados, demandará ambiciosas acciones de capacitación. Y la adaptación creativa de las soluciones europeas a la idiosincrasia institucional patria exigirá dosis no desdeñables de finura jurídica e inventiva reglamentaria.

Aun así, el dictado fundamental que aporta la experiencia europea se revela meridianamente translúcido: toda estrategia de incorporación de la inteligencia artificial a la impartición de justicia debe imprescindiblemente acompañarse de resortes normativos y organizativos que garanticen su supervisión efectiva, su transparencia plena y su plena sujeción a los valores del Estado de Derecho.

3.3.3. Estándares Técnicos y Operativos

Otra de las lecciones cardinales que se colige del análisis del marco regulatorio europeo en materia de inteligencia artificial aplicada a la administración de justicia es la perentoria necesidad de complementar los principios y las garantías de alto nivel con pautas y estándares técnicos y operativos que definan, en términos tangibles y mensurables, cómo deben diseñarse, desarrollarse y desplegarse los sistemas inteligentes para asegurar su alineamiento con los postulados del Estado de Derecho y con las exigencias del debido proceso.

En efecto, los principios de respeto a los derechos fundamentales, no discriminación, calidad y seguridad, transparencia y control humano que vertebran instrumentos como la Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales no se agotan en meras proclamas retóricas, sino que irradian un denso haz de implicaciones prácticas que el legislador comunitario ha sabido traducir en estándares operacionales concretos a lo largo del articulado del Reglamento sobre Inteligencia Artificial (EU AI Act).

Estos estándares, que trazan una verdadera hoja de ruta para la implementación constitucionalmente adecuada de la inteligencia artificial en contextos jurídicamente sensibles como la función jurisdiccional, emergen como uno de los aspectos más valiosos y exportables del acervo normativo europeo, proveyendo orientaciones detalladas que pueden y deben servir de inspiración y referencia para el incipiente proceso de regulación de estas tecnologías disruptivas en el ordenamiento costarricense.

Un primer conjunto de estándares técnicos que el AI Act delinea con singular claridad son los relativos a la calidad y seguridad de los datos utilizados para el entrenamiento, la validación y el funcionamiento de los sistemas inteligentes en el ámbito judicial. Consciente de que los algoritmos son, en última instancia, un reflejo de los datos con los que se alimentan y de que eventuales sesgos, errores o lagunas, en esos datos, pueden traducirse en decisiones erradas o discriminatorias, el legislador europeo ha consagrado un capítulo específico a la "gobernanza de datos" (art. 10), estableciendo que los datasets empleados en sistemas de alto riesgo -como serían, por definición, los aplicados en la administración de justicia- deben ser "pertinentes, representativos, exentos de errores y completos".

Esto implica que el desarrollo de cualquier modelo inteligente orientado a tareas judiciales debe partir de una selección cuidadosa y rigurosa de los datos de entrenamiento, asegurando que estos sean relevantes y apropiados para el contexto y la finalidad del sistema, que abarquen de forma equilibrada y no sesgada el espectro de casos y situaciones sobre los que el algoritmo deberá operar y que estén sujetos a procesos continuos de limpieza, actualización y verificación para evitar imprecisiones o lagunas que pudieran viciar los resultados. Paralelamente, el artículo 15 del Reglamento exige que los sistemas de IA de alto riesgo alcancen niveles apropiados de "exactitud, robustez y ciberseguridad", lo que se traduce en la necesidad de implementar salvaguardas técnicas que minimicen los errores y las vulnerabilidades, además de que permitan detectar y corregir eventuales desviaciones en el rendimiento algorítmico.

Este énfasis en la calidad y seguridad de los datos como precondición para el despliegue legítimo de la IA en sede judicial no es, en modo alguno, baladí. De la fiabilidad e integridad de la información con la que se entrena a los sistemas inteligentes depende, en buena medida, la corrección jurídica de las decisiones o recomendaciones que estos puedan generar. Datos

incompletos, desactualizados o sesgados conducirán irremediablemente a resultados espurios, socavando la equidad procesal y la igualdad de armas. De ahí que cualquier iniciativa de incorporación de herramientas de IA en la justicia costarricense deba acompañarse de protocolos estrictos de gobernanza de datos, auditables por una autoridad independiente, que garanticen el cumplimiento de los estándares reseñados.

Otro pilar fundamental del edificio normativo europeo que ofrece valiosas lecciones para nuestro país son los estándares de transparencia y explicabilidad algorítmica. Consciente de la opacidad consustancial a los modelos de aprendizaje automático y de los peligros que entrañaría una "justicia automatizada" ininteligible, el AI Act consagra un robusto régimen de obligaciones en materia de transparencia (art. 13), trazabilidad (art. 12) y supervisión humana (art. 14) de los sistemas expertos.

Así, los proveedores de IA judicial deberán suministrar información clara, completa y comprensible sobre las características, capacidades y limitaciones de sus sistemas, incluyendo su finalidad, nivel de precisión y principales métricas de rendimiento. Deberán, asimismo, implementar registros automáticos (logs) que permitan seguir y reconstruir el funcionamiento del algoritmo a lo largo de todo su ciclo de vida, posibilitando así la detección temprana de eventuales fallas o desviaciones. Y, sobre todo, tendrán que diseñar sus productos de manera que posibiliten una supervisión humana efectiva, dotando a los operadores jurídicos -jueces, letrados, fiscales- de herramientas de control e interpretación que les permitan entender, auditar e, incluso, desafiar las recomendaciones o predicciones de la máquina.

Este último punto reviste una importancia mayúscula, pues engarza directamente con la independencia judicial como principio estructural del Estado de Derecho. Los estándares de transparencia y explicabilidad buscan conjurar el riesgo de una abdicación de la función jurisdiccional en "oráculos computacionales" inescrutables, preservando la capacidad de los juzgadores de valorar críticamente y, en su caso, apartarse motivadamente de la sugerencia del algoritmo. En este sentido, dichos estándares no son sino la traducción tecnológica de la garantía institucional del "juez natural" predeterminado por la ley, monopolizador indeclinable del acto de juzgar y hacer ejecutar lo juzgado.

Su plena asimilación por el ordenamiento costarricense requerirá no solo de reformas legislativas que expliciten las obligaciones de transparencia y trazabilidad algorítmicas, sino, también, y, sobre todo, de inversiones decididas en la formación digital especializada de los cuadros judiciales. Solo una judicatura tecnológicamente alfabetizada, dotada de las competencias necesarias para comprender y evaluar críticamente el funcionamiento de las herramientas inteligentes a su servicio, podrá evitar el doble escollo de la tecnofobia paralizante y de la aquiescencia acrítica a las decisiones automatizadas.

Un tercer vector de estándares que emerge con nitidez del marco europeo es el atinente a la evaluación y gestión continua de los riesgos que la implementación de IA puede aparejar para derechos y bienes jurídicos fundamentales. Lejos de concebir la introducción de sistemas inteligentes en la justicia como un proceso unidireccional o aproblematífico, el Reglamento comunitario parte de una aproximación realista y cautelar, exigiendo a desarrolladores y operadores un escrutinio permanente de los peligros que, para valores como la igualdad, el debido proceso o la presunción de inocencia, podrían derivarse de un uso inadecuado o defectuoso de los algoritmos.

En este sentido, el artículo 9 del AI Act preceptúa que todo sistema de IA de alto riesgo deberá someterse a un proceso iterativo de gestión de riesgos que abarque la identificación, el análisis, la estimación, la evaluación, la mitigación y la información de los riesgos potenciales. Este proceso deberá llevar aparejado medidas específicas de control del riesgo, como la incorporación de salvaguardas de diseño y defectos, la realización de pruebas de rendimiento *ex ante* y *ex post*, y el suministro a los usuarios de advertencias claras sobre situaciones que puedan dar lugar a riesgos residuales.

Trasladado al contexto judicial, este estándar se traduciría en la exigencia de que toda iniciativa de IA en el seno de la administración de justicia vaya precedida y acompañada de rigurosas evaluaciones de impacto en los derechos fundamentales (EIPD), inspiradas en las populares 'Data Protection Impact Assessments' prevenidas por el Reglamento General de Protección de Datos para tratamientos intensivos de información personal. Estas EIPD deberán documentar pormenorizadamente los potenciales riesgos del sistema para garantías procesales

básicas, como la igualdad de armas, el derecho de defensa, la presunción de inocencia o la tutela judicial efectiva, así como detallar las medidas adoptadas para prevenir o minimizar tales riesgos.

La realización de este ejercicio predictivo y la ejecución subsiguiente de los planes de mitigación y seguimiento no pueden concebirse como un mero formalismo burocrático, sino que han de incardinarse en el corazón de la gobernanza de la IA judicial como precondition de legitimidad y aceptación social. Solo en la medida en que los justiciables y la ciudadanía en general perciban que la incorporación de estas tecnologías disruptivas en los procesos de decisión jurisdiccional va acompañada de análisis de riesgo diligentes y auditables, generadores de salvaguardas tangibles y eficaces, podrá generarse la imprescindible confianza en una "justicia aumentada", capaz de conjugar los beneficios de la eficiencia computacional con la preservación del rostro humano de la judicatura.

Una ulterior dimensión crucial del marco europeo que debe informar el diseño regulatorio costarricense son los mecanismos de evaluación y certificación *ex ante* de los sistemas de IA destinados a emplearse en la administración de justicia. Consciente del altísimo impacto que un algoritmo defectuoso o sesgado podría tener sobre los derechos fundamentales de los justiciables, el legislador comunitario ha sometido las aplicaciones de IA catalogadas de alto riesgo -entre las que figuran expresamente las utilizadas para asistir a las autoridades judiciales- a un exigente régimen de valoración de conformidad previa por parte de organismos independientes acreditados.

Este enfoque preventivo, tributario del principio de precaución que informa toda la estrategia europea de IA, se materializa en la práctica en la exigencia de que todo sistema algorítmico llamado a desplegarse en sede judicial se someta a una batería exhaustiva de pruebas y controles destinados a verificar su fiabilidad, su transparencia, su robustez frente a errores y ciberataques y, sobre todo, su alineamiento con los requisitos legales sustantivos, antes de recibir autorización para operar. El incumplimiento de este trámite preceptivo, además de acarrear cuantiosas sanciones, privaría de toda validez jurídica a las actuaciones practicadas con auxilio de una IA no evaluada o certificada.

Esta apuesta por la verificación *ex ante*, sin duda más onerosa en términos de tiempo y recursos que el mero control judicial *ex post*, revela una comprensión cabal de los formidables

desafíos inherentes a la introducción de sistemas inteligentes en procesos de alto impacto. Desafíos que, en un ámbito tan crucial para la paz social como la administración de justicia, no pueden fiarse al albur de una corrección *a posteriori*, una vez consumada la conculcación de derechos. De ahí la imperiosa necesidad de arbitrar en el ordenamiento jurídico costarricense cauces normativos y procedimentales que instauren, con las necesarias adaptaciones al contexto local, un sistema robusto de escrutinio y homologación preventiva de las soluciones de IA que aspiren a adentrarse en el sacro recinto de los tribunales.

Un último aspecto de la experiencia europea que no puede soslayarse en la construcción de un marco costarricense de estándares para la IA judicial es la imperiosa necesidad de apostar decididamente por la formación y capacitación especializada de los distintos operadores del sistema de justicia en las implicaciones y desafíos de las tecnologías inteligentes. Esta exigencia, explícitamente consagrada en el artículo 14(4) del AI Act y en los principios de la Carta Ética, parte de la constatación de que ninguna supervisión humana de los algoritmos puede reputarse efectiva si quienes han de ejercerla carecen de los conocimientos y las destrezas necesarios para comprender su funcionamiento e implicaciones.

Pretender que jueces, fiscales, letrados o funcionarios judiciales puedan auditar el desempeño de unos sistemas cuya lógica interna les resulta opaca o inaccesible constituiría un ejercicio vano, abocado a la aquiescencia acrítica o a la desconfianza indiscriminada. De ahí la importancia de que toda estrategia de incorporación de IA en la administración de justicia vaya de la mano de planes ambiciosos de alfabetización digital de la judicatura, con itinerarios formativos que abarquen desde los fundamentos técnicos de los algoritmos de aprendizaje automático hasta sus proyecciones jurídicas en ámbitos como la protección de datos, la equidad procesal o la independencia judicial.

Esta formación, lejos de concebirse como un complemento accesorio o voluntarista, debe integrarse en el núcleo de la carrera judicial y del estatuto profesional de los servidores de la justicia, articulándose a través de programas acreditados, de carácter continuo y progresivo. Programas cuyo diseño y ejecución requerirán de la colaboración estrecha entre la Escuela Judicial, los centros universitarios y los colegios profesionales, en un esfuerzo mancomunado por

dotar a los garantes de la tutela judicial efectiva de las competencias necesarias para afrontar los retos de una jurisdicción crecientemente tecnificada.

Concluyendo, el análisis acometido en el presente apartado ha puesto de relieve que los estándares técnicos y operativos emanados del marco regulatorio europeo en materia de inteligencia artificial ofrecen orientaciones valiosísimas para encauzar la implementación de estas tecnologías disruptivas en la administración de justicia costarricense. Desde las exigencias de calidad y gobernanza de los datos empleados en el entrenamiento de los algoritmos hasta los requerimientos de transparencia y explicabilidad de su funcionamiento, pasando por la necesidad de una gestión proactiva de los riesgos, un régimen robusto de certificación previa y una apuesta sin ambages por la capacitación de los cuadros judiciales, el acervo comunitario dibuja una hoja de ruta pionera para un despliegue de la IA en sede judicial que optimice sus potencialidades al tiempo que salvaguarda los derechos fundamentales.

3.3.4.- Reflexiones de Cierre

La construcción del Régimen Europeo sobre IA Judicial no ha sido fruto de una improvisación apresurada, sino de un prolongado ejercicio de equilibrio. Por un lado, existe la pulsión por aprovechar las ventajas objetivas que la automatización e inteligencia artificial ofrecen a la administración de justicia, especialmente ante la acumulación de expedientes y la necesidad de dotar de celeridad al sistema. Por el otro, se mantiene la conciencia de que el Derecho no se reduce a un cálculo numérico, sino que descansa en la ponderación prudencial e individualizada de las circunstancias, en la preservación de la autonomía personal y en la prohibición de que un mecanismo estadístico reemplace la dimensión valorativa y humana de la judicatura. Esa misma dialéctica subyace en las directrices europeas y se concreta en los instrumentos normativos.

Es cierto que, desde una óptica estrictamente tecnológica, la IA se perfila como una herramienta capaz de racionalizar la toma de decisiones, detectando patrones y anticipando desenlaces con un grado de precisión llamativo en determinados contextos. Sin embargo, la mente jurídica europea ha enfatizado que la ratio de la jurisdicción no se agota en la previsión estadística ni en la lógica binaria de la máquina: la equidad, el matiz, la compasión, la dimensionalidad ético-nORMATIVA son elementos que la programación algorítmica difícilmente reproduce de manera

integral. Por ello, la prohibición de la automatización absoluta y la instauración de la intervención humana se interpretan como barreras infranqueables contra la deshumanización del acto de impartir justicia.

La experiencia europea puede servir de catalizador para la adopción de un marco autóctono en cualquier otro país que siga un paradigma constitucional similar, en el sentido de ponderar la libertad y la dignidad en el primer plano. El AI Act, junto con la Carta Ética, no se limita a la abstracción, sino que regula con gran detalle cuestiones prácticas: auditorías, documentación, sistemas de conformidad, normas para la clasificación de usos en bajo o alto riesgo, obligaciones de transparencia para el desarrollador, etc. Esta minuciosidad trasluce el convencimiento de que los principios, para tener una eficacia real, deben irrigar la praxis y permitir mecanismos de verificación. Sin una capa formalizadora de ese talante, toda apelación a la dignidad o a la centralidad humana podría desvanecerse en la inoperancia.

Si bien en la Unión Europea el proceso de implementación del AI Act experimente enmiendas o afinaciones, el conjunto vertebral de su arquitectura —enfoque basado en el riesgo, principios de supervisión humana, no discriminación, transparencia algorítmica— se mantiene inalterado y se presenta como la concreción más avanzada de un proyecto regulatorio global en materia de inteligencia artificial. Su influencia, de manera implícita o explícita, se extiende ya a otros continentes, al perfilarse como un estándar sumamente elevado y abarcador.

La reflexión final se impone: la regulación de la IA en la órbita judicial, si desea amoldarse a un talante garantista, debe reproducir ese grado de exigencia y finura conceptual. El diseño legal no puede ser una mera recolección de buenas intenciones; demanda que se aborden, con la seriedad debida, las implicaciones procesales y éticas de la automatización y que se forjen disposiciones lo bastante específicas para orquestar un control efectivo. El ejemplo europeo, sin ser un dogma, ilustra de modo amplio la factibilidad de avanzar hacia un régimen normativo que preserve la confianza de la población en el carácter humano e imparcial de la justicia, a la vez que abre espacios a la eficiencia e innovación que la tecnología hace posible.

Por ende, la inserción de la IA en la justicia no debe verse como un mero “programa de digitalización”, sino como una relectura de la función jurisdiccional a través del prisma de la

revolución tecnológica, siempre con la impronta humanista que se reconoce en las constituciones democráticas. Esa relectura, para ser fructífera, clama por un marco jurídico que combine la convergencia filosófica (entre IA y derechos humanos), la sofisticación técnica en la regulación y la existencia de una gobernanza institucional sólida. Con todo, la senda emprendida por la Unión Europea constituye un testimonio plausible de que sí es posible conciliar las promesas de la IA —en términos de celeridad, uniformidad, gestión inteligente— con la protección robusta de la persona, sin caer en la desconfianza radical ni en el apresurado automatismo. El paradigma de “justicia algorítmica” que promueve el ordenamiento comunitario se erige, así, en un baluarte de garantías y, al mismo tiempo, en un estímulo para la creatividad regulatoria.

La comprensión de este modelo y su asimilación reflexiva en cada país que pretenda regular la IA judicial debería conducir a la adopción de soluciones jurídicas que, sin repetir ciegamente los artículos europeos, se alineen con su espíritu: la exigencia de que la tecnología esté al servicio de los valores constitucionales y no viceversa. Desde tal premisa, el examen de la experiencia europea nos muestra una arquitectura robusta para trazar ese camino y confiere al legislador, a la judicatura y a la sociedad las pautas orientadoras para encarar una transición digital que sí se maneja con tino y prudencia.

3.4. Propuesta de Implementación

3.4.1. Reformas Legislativas Necesarias

➤ Reformas constitucionales

Análisis de necesidad: la Constitución Política de Costa Rica de 1949 no prevé expresamente el uso de inteligencia artificial en funciones judiciales, pero sí consagra principios fundamentales que deben respetarse. En particular, el artículo 41 garantiza a toda persona “**justicia pronta y cumplida, sin denegación, en estricta conformidad con las leyes**”. Asimismo, el artículo 154 afirma la independencia judicial, estableciendo que “**El Poder Judicial sólo está sometido a la Constitución y a la ley**”. Cualquier uso de modelos de lenguaje (LLMs) en la justicia debe adecuarse a estos principios de acceso a la justicia, debido proceso e independencia.

Possible enmienda: no parece **imprescindible** una reforma constitucional inmediata, siempre que los LLMs actúen como herramientas de apoyo y no reemplacen la función decisoria humana. El **modelo regulatorio europeo** proporciona un referente: el reciente Reglamento de IA de la UE (2024/1689) clasifica como “*alto riesgo*” los sistemas de IA para la administración de justicia, subrayando que **la decisión final debe seguir en manos humanas**. Esta orientación sugiere que la Constitución costarricense, tal como está, ya exigiría ese control humano. No obstante, podría valorarse a futuro una reforma **interpretativa** que explice el derecho a no ser sujeto a decisiones judiciales exclusivamente automatizadas, en línea con garantías internacionales (similar al art. 22 del GDPR europeo que prohíbe decisiones totalmente automatizadas sin revisión humana). Dicha aclaración fortalecería la tutela judicial efectiva y brindaría seguridad jurídica sobre los límites del uso de IA en tribunales.

Fundamentación: una enmienda constitucional podría fundamentarse en la necesidad de **resguardar derechos fundamentales** ante nuevas tecnologías. Por ejemplo, garantizar explícitamente el derecho a un juez **imparcial y humano** en decisiones sustantivas se alinea con el derecho a un juez imparcial ya reconocido. Además, reforzaría el principio pro persona, asegurando que la IA sea un medio para mejorar la pronta justicia sin menoscabar las garantías del debido proceso.

Impacto y desafíos: la ventaja de una reforma constitucional sería brindar un marco supremo claro (ej. estableciendo que las personas tienen derecho a que la IA judicial sea transparente, y a recurrir cualquier decisión automatizada). Sin embargo, modificar la Constitución es complejo: requiere amplios consensos políticos y sociales. Un riesgo es **sobre-regular** tecnológicamente la Carta Magna; podría preferirse esperar a ver cómo evoluciona la implementación práctica y, mientras tanto, confiar en interpretaciones constitucionales dinámicas vía la Sala Constitucional (que podría fijar criterios sobre IA y debido proceso sin necesidad de reforma inmediata). Por ahora, la **viabilidad práctica** sugiere centrar esfuerzos en leyes y reglamentos, usando los preceptos existentes (art. 41 sobre justicia pronta y art. 39 sobre derecho de defensa) como brújula para no violentar derechos.

➤ Modificaciones a la Legislación Primaria

Identificación de leyes a reformar: varias leyes ordinarias requieren actualización para incorporar el uso seguro de LLMs en el Poder Judicial. Una fundamental es la **Ley Orgánica del Poder Judicial** (N.º 7333), que define la estructura y atribuciones judiciales. También las leyes procesales (Código Procesal Civil, Penal, Contencioso) y sectoriales podrían necesitar ajustes. Además, la **Ley 8968 de Protección de la Persona frente al Tratamiento de sus Datos Personales** ya impone obligaciones relevantes, como la despersonalización de sentencias y debe coordinarse con cualquier uso de datos judiciales en IA.

Propuestas Normativas Específicas:

- **Nueva Ley Marco sobre IA:** la regulación del uso de inteligencia artificial en el ámbito judicial requiere, como presupuesto fundamental, la promulgación de una ley marco sobre IA que establezca los principios rectores y el régimen general aplicable a los sistemas automatizados de decisión en Costa Rica. Dicha normativa, siguiendo las tendencias regulatorias más avanzadas como el AI Act europeo, debería adoptar un enfoque basado en riesgos que permita categorizar adecuadamente los diversos usos de la IA según su impacto potencial en los derechos fundamentales y el orden público. Desde este paradigma, las aplicaciones de IA en la administración de justicia quedarían naturalmente clasificadas como de "alto riesgo", lo cual activaría automáticamente un régimen reforzado de salvaguardas, incluyendo evaluaciones de impacto ex ante, requisitos elevados de transparencia algorítmica, estándares de explicabilidad y mecanismos de supervisión humana efectiva. Esta aproximación presenta ventajas significativas desde la perspectiva de economía legislativa y viabilidad política, pues permitiría abordar las especificidades del entorno judicial dentro de un marco normativo más amplio, sin necesidad de promover inicialmente una legislación sectorial específica. Consecuentemente, las iniciativas legislativas que se desarrolleen en la Asamblea Legislativa podrían integrar disposiciones que, además de regular los aspectos comerciales y productivos de la IA, contemplen expresamente las particularidades de su implementación en la función jurisdiccional del Estado y establezcan garantías robustas para la protección de los derechos fundamentales frente a la automatización de procesos judiciales.

- **Reforma a la Ley Orgánica del PJ:** incorporar un artículo que autorice expresamente al Poder Judicial a emplear herramientas de IA como apoyo en sus funciones, **bajo control de autoridades judiciales competentes**. Se podría disponer que la Corte Suprema de Justicia cree un órgano técnico asesor en IA y que toda implementación respete principios de **imparcialidad y no discriminación**. Además, podría aclarar que el uso de IA no vulnera la independencia judicial (art. 154 Const.), siempre que las recomendaciones tecnológicas no sean impuestas externamente sino adoptadas por decisión de los jueces.
- **Adecuación de legislación procesal:** modificar códigos procesales para integrar la IA en el procedimiento. Por ejemplo, en el Código Procesal Civil, adicionar un artículo que permita al juez auxiliarse de herramientas tecnológicas para análisis de jurisprudencia o redactar proyectos de resolución, **sin afectar el derecho de las partes a conocer y controvertir** cualquier información relevante aportada por dichas herramientas. En materia penal, regular el eventual uso de sistemas de apoyo a decisiones (p. ej. evaluación de riesgo de reiteración delictiva mediante IA) garantizando al imputado **derecho a impugnar la base y resultados** de esas evaluaciones. También podría preverse que, si un juez usa un informe o resumen generado por IA para fundamentar su sentencia, dicho informe pase a ser parte del expediente disponible a las partes, garantizando contradicción y transparencia.

Fundamentación jurídica: estas reformas se apoyan en la necesidad de actualizar el ordenamiento a los avances tecnológicos para **hacer efectivo el artículo 41 constitucional** (justicia pronta y cumplida). El uso responsable de IA puede acelerar procesos y mejorar la eficiencia sin comprometer derechos, siempre que la ley imponga salvaguardas. La **Carta Ética Europea de la CEPEJ (2018)** ya identificó principios clave que deben guiar cualquier regulación: respeto a derechos fundamentales, no discriminación, calidad y seguridad, transparencia/imparcialidad, y control humano. Incorporar estos principios en las leyes costarricenses garantizará alineación con estándares reconocidos internacionalmente. Por ejemplo, la **transparencia** algorítmica debe plasmarse en obligaciones legales de explicar a las partes, en términos comprensibles, el rol que jugó la IA en un proceso (análogo al principio de “**explicabilidad**” derivado del debido proceso). Asimismo, la **responsabilidad algorítmica** exige

bases jurídicas: las leyes deben clarificar quién responde ante un error de la IA – previsiblemente, el Estado/Poder Judicial, con posibilidad de repetición al proveedor si hubo falla técnica.

Impacto previsto: las reformas legales proporcionarían un **marco de certidumbre** para la incorporación de LLMs. Con normas claras, los jueces y funcionarios tendrían guía sobre hasta dónde pueden apoyarse en IA (por ejemplo, usar un chatbot para redactar un borrador de sentencia no sería ilegal si la ley lo faculta y regula). Esto impulsará la innovación responsable, evitando tanto la parálisis por temor legal como el uso indiscriminado sin controles. Se espera una mejora en la celeridad de trámites (reducción de la mora judicial) y en la consistencia de las resoluciones, al facilitar acceso rápido a legislación y jurisprudencia relevante mediante IA.

Desafíos de implementación: lograr aprobación legislativa puede ser lento; habrá que concientizar a los diputados sobre la importancia del tema. También, redactar legislación técnicamente sólida en un campo tan novedoso es complejo: deberá involucrarse a expertos en derecho tecnológico. Otro desafío es **no sobrerreglamentar**: fijar principios y lineamientos flexibles, más que detalles técnicos que queden obsoletos. Por ello, muchas disposiciones técnicas podrían delegarse a **normativa secundaria** (reglamentos de la Corte) para poder ajustarlas dinámicamente.

➤ Requerimientos de Normativa Secundaria

Necesidad de regulación interna: además de las leyes formales, el Poder Judicial deberá dictar normativa secundaria (reglamentos, acuerdos de Corte Plena, circulares) para detallar la implementación práctica de los LLMs en los despachos judiciales. Esta normativa, de rango inferior, es más ágil de modificar y puede abarcar aspectos operativos y éticos diarios. En este sentido, resulta imprescindible que dicha normativa se articule en coherencia con el actual "Reglamento del Gobierno, de la Gestión y del uso de los servicios tecnológicos del Poder Judicial" (Reglamento 56-A del 27 de noviembre de 2023), el cual establece las bases para la utilización, seguridad y control de las tecnologías de la información dentro de la institución.

Propuestas específicas:

- **Reglamento sobre el Uso de IA en la Administración de Justicia:** emitido por la Corte Suprema o el Consejo Superior, estableciendo procedimientos y límites para la utilización de IA. Por ejemplo, un reglamento podría obligar que “*cualquier recomendación o producto generado por un sistema de IA debe ser revisado y aprobado por un funcionario judicial antes de incorporarse a una resolución o actuación procesal*”, consagrando el principio de “**bajo control del usuario (juez)**”. También definiría qué tipos de asuntos o tareas son aptos para apoyo de IA (p. ej., clasificación de documentos, búsqueda de jurisprudencia, elaboración de borradores en asuntos de baja complejidad) y cuáles quedan excluidos (p. ej., decisiones en materia penal que afecten directamente la libertad, donde se requerirá especial cautela). Este reglamento debe reflejar los cinco principios de la CEPEJ antes mencionados, traduciéndolos a obligaciones concretas para los operadores judiciales,
- **Integración de principios rectores:** la normativa secundaria debe incorporar expresamente los principios rectores establecidos en el reglamento (seguridad de la información, privacidad, transparencia, confidencialidad y responsabilidad). Estos principios deberán aplicarse a toda implementación de LLMs, garantizando que las soluciones tecnológicas cumplan con los estándares de protección y seguridad definidos (véase Artículo 4 del reglamento),
- **Lineamientos éticos para jueces y funcionarios:** Actualizar el **Código de Ética Judicial** o emitir lineamientos específicos que orienten al personal en el uso responsable de IA. Incluir deberes como: verificar la exactitud de la información aportada por un LLM, no divulgar datos confidenciales a sistemas no autorizados, prevenir sesgos (por ejemplo, corroborar que la IA no esté dando recomendaciones parcializadas contra ciertos grupos). Estos lineamientos pueden inspirarse en la **Carta Ética CEPEJ**, enfatizando valores como transparencia y rendición de cuentas. Un ejemplo práctico: si un juez utiliza un resumen de expediente generado por IA, éticamente debería revelar en el texto que se usó esa herramienta, para mantener transparencia frente a las partes,
- **Protocolos técnicos y de seguridad:** la Dirección de Tecnología del Poder Judicial, con aval de la Corte, debería emitir protocolos sobre integración de IA en los sistemas

judiciales. Por ejemplo, políticas de **gestión de datos** (asegurar anonimización de datos personales, conforme la Ley 8968, antes de usarlos para entrenar modelos) **ciberseguridad** (requisitos de encriptación, control de accesos para evitar injerencias externas en los algoritmos) y **calidad del software** (pruebas periódicas de precisión de los modelos, actualizaciones). Asimismo, procedimientos de **respuesta a fallos**: qué hacer si un sistema de IA presenta errores sistemáticos,

- **Mecanismos de supervisión y auditoría interna:** reglamentar la creación de comités de supervisión incluyendo cómo se reportarán los resultados de auditorías algorítmicas y qué medidas correctivas se tomarán ante hallazgos (p. ej., si una auditoría detecta sesgo racial en las recomendaciones de un LLM para condenas penales, el reglamento debe obligar a suspender el uso de ese modelo hasta corregirlo, en resguardo del principio de no discriminación),
- **Vinculación con la DTIC y la CGTIC:** toda iniciativa de integración de LLMs deberá contar con el visto bueno de la Dirección de Tecnologías de Información y Comunicaciones (DTIC) y ser evaluada y avalada por la Comisión Gerencial de Tecnologías de Información y Comunicaciones (CGTIC), conforme a lo establecido en los Artículos 23, 26 y 36 del reglamento. Esto asegurará que los proyectos se desarrollen dentro del marco estratégico institucional, respetando la lista de software autorizado y la arquitectura tecnológica existente (véase Artículo 35),
- **Procedimientos para actualizaciones y desarrollos tecnológicos:** la normativa secundaria deberá establecer un procedimiento específico para la integración, actualización y mantenimiento de los módulos de IA, de modo que cualquier cambio en el sistema se efectúe siguiendo las metodologías y los lineamientos definidos por la DTIC. Esto incluye:
 - a) La obligatoriedad de utilizar únicamente software que figure en la lista autorizada, garantizando la conformidad con las políticas de derechos de autor y licenciamiento (Artículo 35); b) La definición de un protocolo de integración de nuevos módulos de IA, en coordinación con las áreas de desarrollo tecnológico y bajo la supervisión de la CGTIC.

Fundamentación y buenas prácticas: la **experiencia comparada** muestra la importancia de contar con directrices claras. Organismos internacionales recomiendan instrumentos flexibles: la CEPEJ, por ejemplo, acompañó su Carta Ética con una *lista de verificación práctica* para integrar

los principios en proyectos concretos. Siguiendo esa línea, la normativa interna debe ser detallada y práctica, para guiar a quienes desarrollan o utilizan la IA. Además, recordando que la UE planea exigir que los usuarios de IA de alto riesgo implementen controles y **auditorías externas**, el Poder Judicial costarricense puede adelantarse con regulación propia en ese sentido, aunque no esté obligado por el reglamento europeo, demostrando compromiso proactivo con la excelencia y la rendición de cuentas.

Impacto previsto: la normativa secundaria garantizará una **implementación homogénea** en todos los despachos. Reducirá la incertidumbre de jueces y funcionarios sobre “qué se puede hacer” con IA, fomentando su uso adecuado. También servirá para **tranquilizar a la ciudadanía**: contar con reglamentos públicos sobre IA en justicia aumentará la confianza de los usuarios en que sus casos serán tratados con garantías, aunque intervengan máquinas en el proceso.

Desafíos: habrá que actualizar esta normativa con frecuencia al ritmo de la innovación tecnológica; por tanto, debe preverse un mecanismo de revisión periódica (por ejemplo, que el reglamento sea revisado anualmente por una comisión). Otro desafío es la **capacitación difusa**: todos los operadores deben conocer estas reglas; será necesario divulgarlas ampliamente y quizás certificaciones internas de que cada juez/funcionario ha leído y entendido los protocolos de IA antes de habilitarle el uso de estas herramientas.

No sobra precisar que, si bien la necesidad de un marco legal habilitante resulta indiscutible para dotar de seguridad jurídica a la implementación de la IA judicial, la *naturaleza y densidad* de dicha normativa primaria merecen una consideración estratégica particular. Dada la velocidad exponencial con que evoluciona el campo de la inteligencia artificial (*lex artis* tecnológica), una legislación excesivamente detallada y específica correría el riesgo inminente de obsolescencia normativa. Prescribir en la ley formal requisitos técnicos minuciosos o procedimientos operativos rígidos podría generar un efecto contraproducente, anquilosando la capacidad de adaptación del sistema judicial a futuras innovaciones o a las lecciones aprendidas de la experiencia práctica.

En consecuencia, se recomienda adoptar un enfoque legislativo que priorice la **flexibilidad y la porosidad normativa**. La legislación primaria (ley formal) debería concentrarse en establecer los **principios fundamentales e irrenunciables** (antropocentrismo, supervisión humana, no

discriminación, transparencia básica, responsabilidad), la **clasificación general de riesgos** (identificando explícitamente la IA judicial como de alto riesgo), la **creación del marco institucional de gobernanza** (definiendo la autoridad competente y sus mandatos generales) y las **garantías procesales esenciales** (como el derecho a la revisión humana y a la explicabilidad).

Correlativamente, la **articulación detallada de los estándares técnicos**, los **procedimientos específicos de evaluación de conformidad**, las **metodologías de auditoría algorítmica**, los **protocolos operativos internos** y las **directrices sectoriales concretas** deberían deferirse preferentemente a la **potestad reglamentaria** (normativa secundaria). Esta distribución de competencias permitiría que los aspectos más técnicos y dinámicos fuesen ajustados con mayor agilidad —mediante reglamentos ejecutivos, acuerdos de Corte Plena, circulares de la comisión especializada o directrices de la autoridad competente—, respondiendo así de manera más eficaz a los cambios tecnológicos y a las necesidades emergentes, sin requerir el complejo y dilatado proceso de reforma legislativa para cada adaptación. Este diseño normativo multinivel, centrado en principios robustos en la ley y detalles flexibles en el reglamento, concilia la seguridad jurídica con la necesaria adaptabilidad que impone el ámbito *sub examine*.

3.4.2. Adecuación Institucional

➤ Modificaciones a la Estructura Organizativa

Estado actual: el Poder Judicial, históricamente cimentado en una estructura organizativa tradicional —integrada por la Corte Suprema de Justicia, las salas y diversas judicaturas, complementadas por unidades de apoyo especializadas como la Dirección de Tecnología de la Información (DTI)— ha mostrado, en los últimos años, una marcada inclinación hacia la innovación tecnológica. La instauración de comisiones especializadas, tales como la de Protección de Datos o la Gerencial de Tecnologías de Información y Comunicaciones, ha permitido la implementación de proyectos piloto en materia de inteligencia artificial, destacándose, por ejemplo, la herramienta de despersonalización de sentencias promovida por la Comisión de Protección de Datos. No obstante, la carencia de una unidad dedicada de manera exclusiva a la gobernanza integral de la inteligencia artificial evidencia una laguna en la estructura organizativa, la cual resulta imperiosa subsanar para encarar de forma sistemática y prolongada los desafíos que impone esta tecnología en el ámbito jurisdiccional.

Cambios Propuestos:

• Creación de una Unidad o Comisión de IA Judicial:

Se propone la constitución de un órgano interdisciplinario de carácter permanente, cuya misión fundamental sea la formulación, coordinación y supervisión de las políticas y estrategias dirigidas a la incorporación ética y jurídicamente conforme de la inteligencia artificial en el Poder Judicial. La creación de dicha Comisión encuentra su fundamento en el artículo 13 del Reglamento General de Comisiones de la Corte Suprema de Justicia (según lo establecido en la Circular No. 206-2022, del 12 de diciembre de 2022) y se erige como una respuesta institucional a la necesidad de una orientación estratégica y de largo aliento en un escenario en el que la irrupción de la IA se proyecta como un fenómeno transformador y permanente.

El artículo 11 del citado Reglamento establece una clasificación taxonómica tripartita de las comisiones —permanentes, especiales y especiales con carácter temporal—, la cual responde a la estabilidad y proyección en el tiempo de las funciones encomendadas. Ante la envergadura de los retos que plantea la incorporación de la inteligencia artificial en la administración de justicia, es imperativo que la Comisión se configure como una entidad de carácter **permanente**, asegurando así una respuesta articulada, sistémica y sostenible frente a los cambios disruptivos que se avecinan. Este órgano no solo tendrá la misión de determinar los criterios y estándares técnicos (transparencia, explicabilidad, equidad y respeto irrestricto a los derechos fundamentales) para la implementación de sistemas automatizados, sino, también, la de delimitar los ámbitos jurisdiccionales y las tipologías de casos en los que resulte procedente su aplicación.

La competencia para la creación y conformación de esta Comisión se sustenta en la atribución conferida a la Corte Suprema de Justicia por el artículo 13 del Reglamento General de Comisiones, en consonancia con las potestades generales de gobierno y administración del Poder Judicial recogidas en el artículo 59 de la Ley Orgánica del Poder Judicial. Esta facultad, ejercida con criterio y responsabilidad institucional, demanda una integración equilibrada de perfiles que conjugue la especialización técnica, la pluralidad de perspectivas y la eficiencia deliberativa.

Se sugiere que la Comisión Superior de Regulación y Supervisión de Sistemas de Inteligencia Artificial Judicial (CORSSIA) se integre por **siete miembros**, distribuidos de la siguiente manera:

a) Dos magistrados de la Corte Suprema de Justicia:

- Uno de ellos ejercerá la presidencia, lo que garantizará la máxima legitimidad y una visión estratégica alineada con las políticas institucionales de la judicatura.

b) El director del Departamento de Tecnología de la Información (DTI):

- Su participación es esencial para proporcionar un conocimiento exhaustivo sobre las posibilidades y limitaciones de la infraestructura tecnológica institucional, asegurando que las directrices adoptadas se fundamenten en un rigor técnico indispensable.

c) El director del Departamento de Planificación:

- Su inclusión permite la articulación de los lineamientos en materia de inteligencia artificial dentro de los planes y las estrategias de modernización y optimización del sistema judicial.

d) Un representante de la Escuela Judicial:

- Este componente garantizará la integración de la perspectiva formativa, facilitando la adaptación y capacitación continua de los operadores jurídicos en competencias digitales y tecnológicas.

e) Un profesional en informática con especialización en inteligencia artificial:

- Su experticia será crucial para asegurar que las propuestas y decisiones emanadas de la Comisión se mantengan al nivel del estado del arte en esta disciplina, dotando de rigor científico a sus deliberaciones.

f) Un experto en ética y derechos humanos:

- Su rol será determinante para dotar a la Comisión de una brújula axiológica que prevenga posibles derivas tecnocráticas, garantizando que la implementación de la IA se efectúe siempre en estricto respeto a la dignidad humana y a los principios fundamentales del Estado de Derecho.

La integración de esta estructura especializada y transversal representa una respuesta lógica e instrumental ante la necesidad de modernizar y adaptar el funcionamiento del Poder Judicial a los desafíos del contexto tecnológico actual. La Comisión Superior de Regulación y Supervisión de Sistemas de Inteligencia Artificial Judicial (CORSSIA), en tanto órgano permanente, no solo orientará el proceso de transformación digital, sino que, también, contribuirá a edificar un acervo de conocimiento institucional que permita al Poder Judicial asumir, con solvencia y proactividad, el formidable reto de conciliar la innovación tecnológica con la tutela inquebrantable de los derechos y garantías fundamentales en un Estado Constitucional de Derecho.

Esta propuesta, al conjugar criterios de especialización, transversalidad y operatividad, se erige como la fórmula idónea para garantizar que la adopción de la inteligencia artificial en el ámbito jurisdiccional se efectúe con la debida cautela, rigor técnico y responsabilidad ética, sentando así las bases de un nuevo paradigma en la administración de justicia.

La conformación interdisciplinaria propuesta para la Comisión Superior de Regulación y Supervisión de Sistemas de Inteligencia Artificial Judicial (CORSSIA) busca asegurar una visión holística y experta. Sin embargo, la **celeridad decisoria** constituye un factor crítico en la gobernanza de una tecnología tan dinámica como la IA. Un órgano colegiado con múltiples integrantes de alto nivel jerárquico (varios magistrados, directores de departamento) podría, en la práctica, enfrentar dificultades para adoptar decisiones técnicas u operativas con la agilidad requerida, especialmente si se exige consenso o mayorías calificadas para cada directriz o autorización de pilotos.

En aras de optimizar la **eficiencia deliberativa y ejecutora**, podría valorarse una estructura que, sin sacrificar la representatividad estratégica y la legitimidad institucional, potencie la capacidad de respuesta. Una alternativa a considerar sería **reducir la participación directa de**

magistrados en el pleno de la Comisión —quizás designando a un único magistrado/a representante, con un rol de enlace y garante de la alineación con la política judicial de la Corte Plena— y fortalecer, en cambio, la **presencia de perfiles técnicos y operativos** (expertos en IA, ética, protección de datos, planificación, DTI) con mandatos claros para la formulación de propuestas y la supervisión cotidiana.

Esta reconfiguración no implicaría una merma de la autoridad judicial, sino una delegación funcional razonable de las tareas más técnicas a un cuerpo con mayor especialización y disponibilidad, reservando para la instancia judicial superior (Corte Plena o el magistrado delegado) la aprobación estratégica final y la supervisión de los principios fundamentales. La experiencia comparada sugiere que los órganos ejecutores más expeditos y técnicamente especializados suelen ser más eficaces en la gobernanza de sectores de alta complejidad y rápida evolución. La estructura definitiva deberá, por tanto, ponderar cuidadosamente el equilibrio entre la necesaria representación institucional y la imperativa agilidad operativa que demanda la supervisión de la IA judicial.

- **Unidad Técnica de Inteligencia Artificial**

Complementariamente, se estima necesaria la creación de una Unidad Técnica de Inteligencia Artificial, adscrita a la Dirección de Tecnología de la Información, que tendría a su cargo el desarrollo, supervisión y auditoría constante de los sistemas de IA empleados en la función jurisdiccional. Esta unidad especializada velaría por la calidad, objetividad y pertinencia de los datos utilizados para el entrenamiento de los algoritmos, así como por la trazabilidad y el registro de las decisiones automatizadas. Asimismo, realizaría evaluaciones periódicas para detectar y corregir eventuales sesgos discriminatorios o disfunciones en el desempeño de los sistemas.

Requisitos de implementación: crear estas instancias requiere **recursos humanos especializados** y posiblemente asistencia externa en un inicio. Sería útil apoyarse en convenios con universidades nacionales e internacionales: por ejemplo, invitar a profesores e investigadores a integrar la Comisión de IA, o conformar un **consejo consultivo *ad honorem*** con expertos extranjeros (la experiencia europea o de otros países latinoamericanos piloto podría ser valiosa). Normativamente, bastaría con acuerdos de Corte Suprema para crear estas unidades; no se requiere

ley, ya que la organización interna del PJ se rige en buena medida por decisiones administrativas del propio Poder Judicial (dentro del marco de la Ley Orgánica).

Métricas de éxito: para medir la adecuación organizativa, indicadores podrían ser: (1) **Conformación efectiva** de la Comisión de IA en un plazo determinado (por ejemplo, en el primer año) y número de sesiones o recomendaciones emitidas; (2) Cantidad de **proyectos de IA** lanzados o pilotos coordinados por esta unidad; (3) Nivel de **participación interinstitucional**, medido en convenios firmados o reuniones de coordinación con entes externos; (4) Encuestas internas que midan la **satisfacción/claridad** de jueces y funcionarios respecto a la gobernanza de IA (idealmente mostrando que conocen quién dirige estos esfuerzos y cómo involucrarse). Un éxito cualitativo sería que en pocos años la cultura organizacional acepte la IA como parte de la estructura, de forma similar a como hoy se acepta la existencia de juzgados especializados o comisiones de apoyo.

➤ Requerimientos de Infraestructura Tecnológica

Estado actual: el Poder Judicial costarricense lleva años en un proceso de **digitalización**. Se cuenta con sistemas como el *Sistema de Gestión en Línea* y bases de datos jurisprudenciales (por ej., **NEXUS** para consultas de sentencias). Ya se han desarrollado proyectos piloto que implican infraestructura IA: por ejemplo, en el Juzgado de Cobro de Pérez Zeledón se implementó un sistema que utiliza IA para **clasificar y distribuir escritos** por materia, integrado al software de gestión, lo que mejoró la celeridad procesal y redujo el rezago. Este y otros proyectos utilizan los recursos de servidores del PJ y la infraestructura de la DTI. A continuación, se detallan las necesidades y requerimientos clave para la expansión e integración de los LLMs (Modelos de Lenguaje de Gran Escala) y otras herramientas de IA en el ámbito judicial, incorporando además la estimación de costos y la configuración de hardware/ software sugerida por Matthew Carrigan³²¹ (ingeniero de HuggingFace, empresa puntera en IA de código abierto) para correr localmente el modelo DeepSeek R1, uno de los más avanzados y que **requiere apenas unos USD 6,000** en equipamiento. Esta propuesta resulta sumamente atractiva para el Poder Judicial, pues permite

³²¹ Carrigan, Matthew (@carrigmat). “Complete hardware + software setup for running Deepseek-R1 locally. The actual model, no distillations, and Q8 quantization for full quality. Total cost, \$6,000. All download and part links below:” *Hilo en X*, 28 de enero de 2025, 8:17 a.m. Recuperado de: <https://x.com/carrigmat/status/1884244369907278106>

mantener los datos sensibles bajo total control institucional, sin necesidad de cederlos a terceros o utilizar servicios cerrados (closed-source) con riesgos de confidencialidad.

Necesidades Identificadas:

- **Capacidad de Cómputo y Almacenamiento**
 - **Necesidad de recursos de cómputo**
 - Los LLMs demandan recursos de procesamiento considerables, tanto en CPU como en GPU. Sin embargo, existe la opción de modelos optimizados que pueden ejecutarse íntegramente en CPU si se cuenta con suficiente ancho de banda de memoria y RAM,
 - Ejemplo: El modelo DeepSeek R1, según la estimación de Carrigan, puede correr en un sistema con procesadores AMD EPYC (serie 9004 o 9005) y 768 GB de RAM DDR5, con un rendimiento adecuado para realizar inferencias y tareas de razonamiento avanzado.
 - **Estimación de costos para una solución local**
 - Siguiendo las recomendaciones de Carrigan, se pueden adquirir componentes de mercado con un costo total aproximado de **USD 6,000**, suficientes para correr **DeepSeek R1** (modelo open-source de vanguardia)
 - **Tarjeta madre:** Gigabyte MZ73-LM0 o MZ73-LM1, con 2 sockets EPYC para disponer de 24 canales de DDR5,
 - **Procesadores:** 2 × AMD EPYC (modelo 9004 o 9005; por ejemplo, 9115 o 9015 para reducir costos). Esto equilibra costo y desempeño, pues en LLMs el **cuello de botella** suele ser el ancho de banda de la memoria,
 - **Memoria RAM:** 768 GB en total (24 módulos de 32 GB DDR5-RDIMM), a fin de permitir cargar y ejecutar el modelo sin compromisos de velocidad,
 - **Chasis / Case:** Un gabinete que soporte factor de forma E-ATX/SSI-EEB (por ejemplo, el Phanteks Enthoo Pro 2 Server Edition),

- **Fuente de poder (PSU):** 1000 W, con suficientes cables de alimentación para dos CPUs EPYC (ej. Corsair HX1000i),
 - **Refrigeración / Disipadores:** Para socket SP5 de EPYC, se requieren disipadores compatibles (p. ej., disponibles en eBay o Aliexpress), que luego pueden complementarse con ventiladores silenciosos (Noctua NF-A12x25),
 - **Almacenamiento:** SSD NVMe de al menos 1 TB, que permita cargar con rapidez los ~700 GB del modelo,
 - **Costo estimado:** ~USD 6,000 por todo el hardware + licencias/ajustes de software necesarios.
- La configuración de hardware ejemplificada anteriormente ilustra el tipo de sistema de alto rendimiento que el Poder Judicial necesitaría para embarcarse en la adopción local de modelos de lenguaje de gran escala (LLMs). Equipos con estas características –**procesadores de servidor potentes, memoria RAM muy abundante y almacenamiento rápido**– son capaces de procesar los grandes volúmenes de información inherentes a las tareas judiciales y responder a las solicitudes de los usuarios. Crucialmente, la ejecución de modelos avanzados, como los de código abierto mencionados [*como DeepSeek R1*], en **infraestructura propia** permite operar **sin compartir datos judiciales sensibles con terceros**, un requisito fundamental para salvaguardar la privacidad de los expedientes, garantizar la seguridad de la información y mantener la soberanía institucional sobre los datos.

En términos de capacidad, un servidor individual de estas características puede gestionar **múltiples interacciones simultáneas**, aunque el número exacto dependerá de la complejidad de las consultas y del modelo específico en uso. Sin embargo, para un escenario realista donde **cientos de jueces, letrados y personal administrativo** interactúen con el sistema de IA de forma concurrente (realizando consultas, generando textos, analizando documentos), **será indispensable implementar un clúster compuesto por una cantidad significativa de estos servidores**.

La **escala de dicho clúster**, y por ende la inversión total en hardware, deberá dimensionarse cuidadosamente en función de la carga de usuarios proyectada, los tipos de tareas a realizar y los niveles de rendimiento y redundancia deseados. Un **enfoque modular y escalable** se presenta como el más adecuado: iniciar con un número limitado de servidores para proyectos piloto o fases iniciales, y **expandir la infraestructura progresivamente** a medida que la demanda crezca y se validen los beneficios. Esta flexibilidad permite ajustar la inversión a las necesidades reales y a la capacidad presupuestaria del Poder Judicial.

Finalmente, es imperativo recordar que la inversión inicial en hardware (servidores, memoria, procesadores) representa solo una parte del costo total de propiedad. A esto deben sumarse los **gastos operativos recurrentes**, que incluyen el mantenimiento de los sistemas, el consumo energético y la refrigeración, el licenciamiento de software (si aplica, aunque se prioricen soluciones abiertas), y de manera muy importante, la **inversión sostenida en personal técnico especializado** para administrar, supervisar y mantener esta infraestructura avanzada, así como la **capacitación continua** de los usuarios finales. No obstante, dado que este modelo de implementación local prioriza la **autonomía tecnológica y la máxima confidencialidad de los datos**, se considera **estratégicamente idóneo** para una institución como el Poder Judicial, que maneja información de alta sensibilidad y debe operar bajo estrictos estándares de integridad, transparencia y seguridad.

- **Integración con sistemas existentes:** la infraestructura debe permitir que los módulos de IA se acoplen a los sistemas judiciales actuales. Por ejemplo, integrar un modelo de lenguaje que ayude a redactar resoluciones dentro del mismo entorno donde el juez las elabora (así el juez no tiene que salir a otra aplicación, con los riesgos de seguridad que ello implica). Esto requiere **APIs y desarrollo de software** personalizado. La DTI deberá ampliar o adaptar las plataformas judiciales (como el gestor de casos) para tener funcionalidades “inteligentes”. Se podría adoptar una arquitectura modular, donde componentes de IA (p.ej., motor de procesamiento de lenguaje natural) se comuniquen con el sistema principal mediante interfaces bien definidas,

- **Datos y digitalización:** un elemento crítico es **contar con datos de calidad para entrenar y operar los LLMs**. El Poder Judicial debería consolidar sus bases de datos de jurisprudencia, leyes y expedientes en formato digital utilizable. Es probable que se requiera un gran esfuerzo de **limpieza y marcado de datos**: por ejemplo, etiquetar decisiones por tipo de asunto, resultado, etc., para entrenar modelos que, dados unos hechos, sugieran una solución basada en precedentes. También, asegurar la **despersonalización** (anonimización) de datos sensibles antes de usarlos en entrenamiento, en cumplimiento de la normativa de privacidad. Esto podría suponer ampliar el proyecto de despersonalización existente con mayor capacidad, dado que actualmente se aplica a sentencias finales, pero para IA sería útil anonimizar también expedientes completos,
- **Seguridad informática:** la introducción de IA amplía la **superficie de riesgo** de ciberataques o manipulación. Imaginemos las consecuencias de un ataque que altere un modelo de IA para sesgar sus recomendaciones en favor de cierto resultado. La infraestructura deberá incluir medidas de **seguridad reforzada**: monitoreo constante, auditorías de código, control de versiones (para detectar cambios no autorizados en los algoritmos), y posiblemente **entornos aislados** para pruebas. Además, dado que la IA podría tomar decisiones administrativas (ej. asignar un caso a determinada sala), se deben implementar logs y trazabilidad de cada acción que el sistema hace, de modo que cualquier irregularidad pueda investigarse.

Requisitos de implementación: primero, se debe realizar un **diagnóstico técnico** de brechas: ¿Cuánto almacenamiento y cómputo adicional se necesita? ¿Qué tan actualizados y compatibles son los sistemas actuales con IA? Este diagnóstico guiará un **plan de inversiones**. Es fundamental asegurar presupuesto plurianual, posiblemente reorientando partidas de modernización o buscando financiamiento externo (cooperación internacional, préstamos BID/BM orientados a justicia digital, etc.). En paralelo, la DTI debe **capacitarse** o contratar expertos para diseñar esta ampliación de infraestructura. Un cronograma claro ayudará a priorizar: quizás en corto plazo se adapte infraestructura para un par de pilotos específicos, y en mediano plazo se adquieran sistemas más robustos para escalamiento.

Métricas de éxito: podrían medirse: (1) **Disponibilidad** y tiempos de respuesta de los sistemas de IA (que indiquen si la infraestructura soporta la carga); (2) Número de **incidentes de**

seguridad relacionados con IA (esperando que sea cero, o tendiente a cero gracias a buenas prácticas); (3) Proporción de **datos judiciales digitalizados y listos** para IA (por ejemplo, porcentaje de expedientes históricos incorporados a bases electrónicas); (4) **Índice de integración**: cuántos sistemas de IA están efectivamente integrados a plataformas usadas por jueces vs. cuántos funcionan aislados o en prueba. Un signo de éxito sería que los usuarios reporten que las herramientas de IA se sienten como parte natural del sistema informático judicial, sin fricciones técnicas.

➤ Desarrollo del Recurso Humano

Estado actual: el capital humano del Poder Judicial abarca jueces, letrados, personal administrativo, técnicos de TI, etc. Actualmente, la mayoría no tiene formación especializada en IA. Sin embargo, hay antecedentes positivos: se han realizado **capacitaciones** en tecnologías (por ejemplo, entrenamientos en expediente digital) y recientemente, por impulso internacional, jueces costarricenses participaron en talleres sobre “**Inteligencia Artificial y Estado de Derecho**” organizados por UNESCO y la Corte IDH. En dicho programa (noviembre 2023, San José), más de 40 jueces de la región discutieron beneficios y riesgos de la IA en la justicia, incluyendo sesgos, “cajas negras” y falta de transparencia. Esto indica una **conciencia inicial** y disposición a aprender.

Cambios Propuestos:

- **Capacitación continua y transversal:** institucionalizar programas de formación en IA para **todos los niveles** del Poder Judicial. La **Escuela Judicial** deberá incorporar en su malla curricular cursos sobre tecnología y ética de la IA. Nuevos jueces y funcionarios en inducción deberían recibir nociones de qué son los LLMs, cómo pueden asistir en las tareas judiciales y cuáles son sus limitaciones. Para jueces en ejercicio, organizar seminarios y talleres prácticos: por ejemplo, entrenamientos de cómo usar una herramienta de resumen automático de expedientes, o cómo interpretar un报告 generado por IA. Importante también incluir a **personal administrativo y auxiliares** que puedan verse reubicados en funciones: capacitarlos en supervisión de sistemas (por ejemplo, un asistente legal podría pasar a revisar el output del IA antes de dárselo al juez). Se puede buscar apoyo de

organismos internacionales (CEPEJ, UNESCO, etc.) para estas capacitaciones, dado que ya han mostrado interés en preparar jueces en estos temas,

- **Sensibilización ética:** más allá de lo técnico, formar en la **cultura de la IA responsable**. Impartir charlas sobre los principios éticos (no discriminación, respeto derechos, etc.) para que el personal comprenda la razón de las nuevas reglas. Esto cultivará criterio en los operadores: un juez capacitado sabrá detectar si una recomendación del sistema puede estar sesgada o ser contraria a derechos, y tendrá la confianza de contrariarla si es necesario (evitando sumisión ciega a la máquina). También se deben dar a conocer las experiencias comparadas: estudios de casos de otros países donde la IA fue útil o donde hubo problemas (p.ej., se podría analizar el caso COMPAS de EE.UU. sobre sesgo racial en evaluación de reincidencia, el piloto de “juez digital” en Estonia y los Smart Courts en China).
- **Especialización de personal TI interno:** desarrollar fuertemente el **recurso humano técnico**. Esto implica capacitar a los actuales ingenieros y desarrolladores en técnicas de machine learning, ciencia de datos jurídica y manejo de LLMs, posiblemente mediante cursos avanzados, certificaciones internacionales o estancias cortas en instituciones pioneras. Además, podría requerir **nuevas contrataciones**: data scientists, analistas de datos, expertos en PLN (procesamiento de lenguaje natural) con dominio del idioma español y si es posible, conocimientos jurídicos. Dada la limitación presupuestaria, quizás convenga un enfoque gradual: formar al personal TI existente que tenga aptitud, complementándolo con consultores externos por proyecto mientras se transfiere conocimiento. Incentivar convenios con universidades locales para pasantías de estudiantes de informática o inteligencia artificial en el PJ, creando una cantera de talento futuro,
- **Gestión del cambio organizacional:** implementar planes de **gestión de cambio** para abordar temores o resistencia. Algunos empleados podrían temer que la automatización amenace sus puestos; es clave comunicar que la IA viene a asumir tareas repetitivas y a liberar al personal para tareas de mayor valor, no a sustituir la función jurisdiccional. A tal efecto, se pueden re-definir **perfiles de puestos** gradualmente: por ejemplo, un auxiliar de tribunal que antes transcribía minutas ahora podría convertirse en un “verificador de IA”, encargado de validar los documentos generados automáticamente. Mostrar ejemplos exitosos (como el piloto de Cobro Judicial que **mejoró el rendimiento sin despidos**) ayudará a reducir la ansiedad. También, reconocer y visibilizar logros: premiar a despachos

que adopten exitosamente la nueva tecnología, de modo que se vea como algo positivo en la carrera judicial adaptarse a estas herramientas.

Requisitos de implementación: será necesario destinar **presupuesto** para capacitación (cursos, seminarios). Podría provenir de partidas de capacitación ordinarias, reorientadas a este tema prioritario. Además, elaborar un **plan de capacitación 2025-2030** con metas: por ejemplo, que al cabo de 2 años el 100% de los jueces de primera instancia hayan asistido al menos a un taller de introducción a IA; a 5 años, que exista un diplomado en Jurimetría/IA Judicial en la Escuela Judicial para quienes profundicen. La alta dirección debe impulsar esto con instrucciones claras a las jefaturas de que faciliten la participación del personal en entrenamientos (liberándolos de carga cuando sea necesario).

Métricas de éxito: (1) **Cobertura de capacitación:** porcentaje de funcionarios capacitados en temas de IA; (2) **Nivel de competencia adquirido:** evaluaciones pre/post capacitación que muestren aumento en el conocimiento o en la confianza para usar las herramientas; (3) **Adopción efectiva:** número de usuarios activos de cada herramienta de IA (si muchos no la usan por falta de habilidades, será indicativo de necesidad de más formación); (4) **Cambio de actitud:** mediante encuestas de clima organizacional o entrevistas cualitativas, evaluar la percepción del personal sobre la IA (meta: mayoría la vea como ayuda útil y no como amenaza). También se puede medir si la capacitación reduce errores de uso: por ejemplo, rastrear casos en que un juez tomó una acción inapropiada por mal uso de la herramienta y ver que estos incidentes tiendan a cero con mejor entrenamiento.

➤ Mecanismos de Control y Supervisión

Estado actual: el Poder Judicial cuenta con mecanismos de control tradicionales: auditoría interna, inspección judicial para conducta de jueces, la Sala Constitucional revisando posibles violaciones de derechos (vía recursos de amparo contra actuaciones judiciales) y la propia segunda instancia corrigiendo errores. Sin embargo, no existen aún **mecanismos específicos para la supervisión de sistemas de IA**. Dado que se han implementado proyectos piloto de IA, es probable que informalmente la DTI y las comisiones responsables realicen seguimiento de

resultados, pero no bajo un marco estructurado de auditoría algorítmica o de evaluación de impacto continuo.

Cambios Propuestos:

- **Evaluaciones de impacto y auditorías periódicas:** institucionalizar la práctica de realizar **evaluaciones de impacto algorítmico** antes de la implementación de cada nuevo LLM y auditorías periódicas después. Esto implica conformar equipos mixtos (tecnólogos + juristas + eventualmente observadores externos) que examinen aspectos clave: posibles sesgos en la toma de decisiones de la IA (p. ej., ¿el modelo de lenguaje tiende a favorecer un tipo de argumento legal sobre otro de modo injustificado?), tasas de error, coherencia con cambios legales (un riesgo es que el modelo se quede desactualizado si la ley cambia, por lo que debe verificarse su base de conocimiento). Una auditoría podría, por ejemplo, revisar aleatoriamente 100 recomendaciones de una IA en casos ya resueltos para ver cuántas divergían del resultado correcto; o simular entradas con variaciones (cambiando nombres, género) para detectar discriminación. Los resultados de estas evaluaciones deben elevarse a la Comisión ya planteada y a la Corte, con **planes de mejora** si se hallan problemas. Idealmente, algunos de estos informes podrían hacerse públicos para generar confianza (resguardando información sensible),
- **Supervisión jurisdiccional:** asegurar que las **instancias superiores** estén atentas a cómo la IA pudo influir en casos. Si notan que un juez basó su fallo excesivamente en un reporte de IA sin suficiente fundamento propio, podrían señalarlo en la sentencia de alzada, construyendo así una jurisprudencia sobre el uso correcto de IA. Con el tiempo, estas directrices jurisdiccionales complementarán la normativa escrita. La **Sala Constitucional** también jugará un rol: por ejemplo, podría recibir un amparo de alguien alegando que una decisión fue tomada por “una computadora y no por un juez”. Aunque en teoría eso no debe ocurrir con los controles propuestos, si llegara a pasar, la Sala IV podría fijar estándares de debido proceso aplicables (dando una última capa de garantía). Incorporar estas posibilidades de control jurisdiccional en el diseño (p. ej., facilitando que en los expedientes conste la información necesaria para que un tribunal superior entienda la intervención de la IA),

- **Monitor de rendimiento y calidad:** además de los controles éticos/jurídicos, implementar sistemas automáticos de monitoreo de la **calidad de las decisiones de IA**. Por ejemplo, un dashboard que mida el porcentaje de coincidencia entre las recomendaciones de IA y las decisiones humanas finales (no porque se busque 100 % coincidencia, sino para detectar desviaciones significativas: si la IA constantemente sugiere algo que los jueces rechazan, puede que esté mal calibrada). Otros indicadores: tiempo promedio que ahorra la IA, número de correcciones manuales requeridas, etc. Este monitoreo continuo permitirá calibrar los modelos y demostrar su eficacia o alertar de fallas rápidamente.

Estado actual vs. cambios propuestos: actualmente no hay nada de esto; con los cambios, pasaríamos a una cultura de “**algoritmos bajo escrutinio**” constante, lo cual es esencial para la **responsabilidad algorítmica**. A nivel internacional, se insiste en que la IA en justicia debe ser **auditada regularmente y mejorada continuamente**. El **Reglamento de IA de la UE** exige algo similar para sistemas de alto riesgo (proveedores deben permitir auditorías y registro en base de datos de sistemas). Costa Rica puede voluntariamente adoptar esas prácticas.

Requisitos de implementación: para realizar auditorías, se requerirá **expertise técnico y tiempo**. Una opción es contratar terceros independientes (auditorías externas) para revisiones anuales de los algoritmos más críticos, lo que da imparcialidad. También, capacitar a la Auditoría Interna existente en nociones de auditoría de algoritmos, o crear una pequeña unidad técnica de auditoría de IA dentro del Poder Judicial. Normativamente, se pueden incorporar en reglamentos la obligación de someter los sistemas a estas evaluaciones. Es clave también tener cooperación de los **desarrolladores**: en los contratos con proveedores de IA, estipular que deberán brindar acceso al código o modelo para fines de verificación (respetando propiedad intelectual, se pueden hacer bajo acuerdos de confidencialidad, pero sin acceso sería imposible auditar, lo cual no sería aceptable en sistemas que afectan derechos).

Métricas de éxito: (1) **Número de auditorías realizadas** vs. planificadas (cumplimiento del calendario de evaluaciones); (2) **Recomendaciones emitidas vs. implementadas**: cuántas recomendaciones de mejora surgen de los controles y cuántas se ejecutan efectivamente (buscando una tasa alta de implementación, señal de compromiso de mejora continua); (3) **Incidencias éticas resueltas**: cantidad de reportes recibidos por el comité ético y porcentaje solucionado

satisfactoriamente; (4) **Indicadores de confianza pública:** por ejemplo, encuestas a usuarios o un seguimiento de menciones en medios: ¿la gente percibe que la IA judicial es confiable y está bajo control? Si los mecanismos funcionan, debería evitarse cualquier escándalo o pérdida de confianza, manteniendo la reputación del PJ. Un objetivo podría ser que **ninguna resolución sea anulada** por defectos atribuibles a IA; es decir, que los controles prevengan violaciones de debido proceso antes de que ocurran.

3.4.3. Hoja de Ruta

➤ Objetivos a Corto Plazo (1-2 años)

En los primeros 1-2 años, la prioridad será sentar las bases normativas e institucionales y lanzar proyectos piloto controlados.

Objetivos Principales a Corto Plazo:

- **Marco normativo inicial:** elaborar y promover ante la Asamblea Legislativa las reformas legales identificadas. En el año 1, podría presentarse un proyecto de reforma a la Ley Orgánica del PJ y, dependiendo del ambiente político, apoyar la discusión del proyecto de ley general de IA que abarque la gestión de riesgos propuesta. Paralelamente, la Corte Suprema debe emitir en este periodo al menos un **Acuerdo o Reglamento provisional** sobre el uso de IA en sede judicial, para regular los pilotos en lo que las leyes formales se aprueban.
 - *Hito:* tener aprobado (o en avanzada discusión legislativa) el marco legal primario para finales del segundo año.
- **Creación de instancias y planificación:** constituir la **Comisión Superior de Regulación y Supervisión de Sistemas de Inteligencia Artificial Judicial (CORSSIA)** (u órgano equivalente) dentro de los primeros 6 meses. Esta comisión, una vez formada, debe en los siguientes meses delinear un **Plan Estratégico de IA 2025-2030** del Poder Judicial, que sirva de guía.
 - *Hito:* comisión instalada y Plan Estratégico publicado antes de cumplir 2 años.
- **Pilotos tecnológicos focalizados:** implementar **proyectos piloto** de bajo riesgo que demuestren resultados rápidos. Por ejemplo, en el año 1: extender el piloto de **clasificación**

automática de escritos a 2 o 3 despachos adicionales aparte de Pérez Zeledón, midiendo la reducción de tiempos. Iniciar un piloto de **asistente virtual para el público** (chatbot legal) en alguna materia frecuente (familia o pequeñas causas) para responder preguntas comunes, evaluando su precisión y satisfacción de usuarios. También podría lanzarse un prototipo de **LLM entrenado en jurisprudencia costarricense** para ayudar a jueces a encontrar precedentes (quizá comenzando con un conjunto limitado, como toda la jurisprudencia de Sala Constitucional sobre un tema).

- *Hito:* al menos 2 pilotos iniciados en el año 1, con informes de resultados preliminares en el año 2.
- **Capacitación inicial y sensibilización:** en el corto plazo, realizar **jornadas de capacitación masivas**. Por ejemplo, en el primer año, organizar un ciclo de conferencias con expertos internacionales (aprovechando el interés de UNESCO y la UE en el tema) para magistrados y jueces. También, talleres prácticos para personal administrativo que participará en los pilotos (asegurarse de que los usuarios de los proyectos piloto estén bien formados).
 - *Hito:* 100 % de los jueces de despachos piloto capacitados antes de usar la herramienta; al menos 30 % de los magistrados y jueces a nivel nacional asistieron a alguna charla introductoria en el bienio.
- **Infraestructura y datos:** a corto plazo, inversiones modestas pero cruciales: adquirir servidores o habilitar entornos en la nube para correr los pilotos sin afectar sistemas actuales. Iniciar la **unificación de bases de datos**: por ejemplo, compilar todas las sentencias de 2010-2024 de las Salas en una base de conocimiento consolidada para entrenar al LLM jurídico.
 - *Hito:* disponibilidad de un repositorio unificado de jurisprudencia y legislación en formato utilizable por IA al cabo de 2 años.
- **Gestión del cambio:** en estos primeros años, la comunicación es clave. Lanzar una **campaña interna de difusión** sobre la estrategia de IA, utilizando los canales del PJ (boletines, intranet, reuniones) para explicar objetivos y disipar temores. Establecer un canal de retroalimentación (inbox o foro interno) donde los funcionarios puedan comentar preocupaciones o ideas respecto a la IA, y la Comisión de IA las atienda.

- *Hito:* tener documentadas las preocupaciones comunes levantadas en el año 1 y emitir una FAQ o lineamientos para responderlas.

Evaluación de riesgos (corto plazo): los riesgos iniciales incluyen **resistencia al cambio**, fallos técnicos en pilotos, o retrasos legislativos. Para mitigarlos: la sensibilización y participación temprana del personal mitigará resistencia; empezar con pilotos sencillos y con plan B manual reducirá impacto de fallos; y si la ley tarda, usar normativa interna temporal mantendrá el proceso avanzando. Un riesgo específico es implementar algo sin suficiente preparación y que ocurra un error que desacredite el programa (p. ej., un chatbot que dé información legal errónea públicamente). Manejo: limitar el piloto a entornos controlados, supervisar respuestas antes de permitir publicación y tener mensajes preparados para aclarar que está en prueba. El éxito en corto plazo se medirá en la **instalación**: que existan los cimientos legales, institucionales y tecnológicos para luego escalar.

Criterios de evaluación (corto plazo): al terminar el año 2, la evaluación se centrará en si se cumplieron los hitos descritos. Por ejemplo: ¿Está operando la Comisión de IA y el reglamento interno? ¿Qué resultados cuantitativos muestran los pilotos (ej. X % reducción de tiempo en clasificación de documentos)? ¿El nivel de participación en capacitaciones alcanzó lo previsto? Un informe de logros vs metas deberá presentarse a la Corte Suprema y podría compartirse con la opinión pública para mantener el apoyo.

➤ Metas a Mediano Plazo (3-5 años)

En 3-5 años, se espera pasar de pilotos a **implementación institucional más amplia** y afinar el marco regulatorio con la experiencia obtenida. **Metas a mediano plazo:**

- **Aprobación completa del marco legal:** para este horizonte, idealmente la **ley general de IA** en Costa Rica habrá sido aprobada (posiblemente con un capítulo dedicado a sector justicia) o se habrán introducido las reformas en leyes judiciales clave. También la normativa interna habrá madurado: podrían emitirse versiones actualizadas de reglamentos tras aprender de los primeros años.
 - *Hito:* para el año 5, existencia de un cuerpo legal en vigor que regule IA en la justicia, armonizado con estándares internacionales (por ejemplo, que incorpore

formalmente los principios de transparencia, equidad, etc., tal como la UE los exige y CEPEJ recomienda-

- **Escalamiento de herramientas IA:** las soluciones piloto exitosas deberán escalarse a nivel nacional. Por ejemplo, el sistema de **clasificación automática de escritos** implementado en todos los juzgados de cobro y luego adaptado a otras jurisdicciones (familia, laboral para filtrado de casos, etc.). Un **LLM jurídico entrenado en el ordenamiento costarricense** podría estar a disposición de cada juez como asistente personal: para el año 5, un juez de cualquier provincia debería poder preguntarle al sistema por jurisprudencia relevante o incluso pedirle un borrador de resolución para editar, con la confianza de que es fiable. Asimismo, se puede introducir IA en áreas administrativas: gestión de agenda, asignación de expedientes a despachos según carga de trabajo, etc., optimizando recursos.
 - *Hito:* a los 5 años, al menos 3 herramientas basadas en IA (ej. asistente de redacción de sentencias, chatbot de atención ciudadana, clasificador de casos) operando en **producción** en la mayoría de oficinas judiciales, con manuales de uso y soporte técnico estable.
- **Mejoras en indicadores judiciales:** se espera ver reflejado el impacto en indicadores medibles de servicio de justicia. Metas como **reducir la duración promedio de los procesos** en cierto porcentaje gracias a la agilización que proporciona la IA, o **bajar la tasa de atraso judicial** en áreas donde se implementó. Por ejemplo, si al inicio la jurisdicción contencioso-administrativa tenía un tiempo promedio de resolución de 3 años, quizá con herramientas de gestión inteligente y priorización (IA que indique cuáles casos llevan mucho tiempo para darles impulso) se logre bajar a 2 años.
 - *Hito:* mejora significativa (10-20 %) en tiempos de trámite o reducción de carga en las materias intervenidas por IA, al cabo de 5 años, según los reportes del Poder Judicial.
- **Consolidación del talento humano:** para el año 5, la meta es tener un **equipo interno consolidado** capaz de continuar el desarrollo y mantenimiento de sistemas de IA. Esto incluye no solo técnicos, sino también jueces “líderes en IA” en cada circuito judicial que actúen como referentes y mentores de sus pares en el uso de las herramientas.

- *Hito:* red de al menos 20 “jueces innovadores” formados, uno por cada gran circuito o materia, y un departamento de IA con personal estable en la DTI. Adicionalmente, incluir módulos de IA en los concursos de ingreso o ascenso de la carrera judicial podría considerarse a esta altura, para motivar que los aspirantes vengan familiarizados con tecnología.
- **Supervisión y mejora continua:** a mediano plazo debe estar funcionando robustamente el **sistema de control**. Las auditorías anuales de los algoritmos se habrán realizado y el proceso de **certificación** o validación interna de cada nueva versión de un modelo será rutinario.
 - *Hito:* publicación de informes anuales de la Comisión de IA mostrando evaluaciones de impacto, incidencias detectadas y cómo se corrigieron, demostrando un ciclo de mejora continua. Por ejemplo, si en año 3 se detectó cierto sesgo en un modelo, para año 4 ese sesgo esté corregido y documentado. También que los órganos de control externo (Sala Constitucional, ARIA si existe) **no hayan tenido que intervenir adversamente**, es decir, no haya sentencias declarando inconstitucional el uso de tal o cual sistema gracias a que internamente se manejaron bien los riesgos.
- **Participación y colaboración internacional:** para el año 5 Costa Rica podría ser un ejemplo regional en justicia digital. Metas podrían ser: organizar una **conferencia internacional** en San José sobre IA y justicia para intercambio de experiencias, o exportar el modelo a otros países centroamericanos. Esto no es un objetivo interno per se, pero indica liderazgo.
 - *Hito:* Costa Rica participando en proyectos colaborativos (por ej., compartiendo datasets anonimizados con iniciativas globales de IA jurídica, o implementando estándares de interoperabilidad con otros países para intercambio de información judicial asistido por IA).

Evaluación de riesgos (mediano plazo): posibles riesgos en esta etapa incluyen: **estancamiento tecnológico** (que tras pilotos no se logre implementar masivamente por falta de presupuesto o voluntad), **problemas de confianza pública** (un error sonado podría hacer que la opinión pública o algún gremio se oponga a seguir), o **desfase regulatorio** (que la tecnología

avance más rápido que la regulación ajustada en corto plazo, requiriendo nuevas reformas a mitad de camino). Para mitigarlos: asegurar financiamiento sostenible, celebrar los éxitos tempranos públicamente para mantener apoyo; mantener la transparencia total (si ocurre un fallo, reconocerlo y corregirlo abiertamente, mostrando la solidez de los controles) y estar listos para ajustar normativas internas rápidamente si la realidad muestra algo distinto de lo esperado. La Comisión de IA deberá estar monitoreando tendencias (por ejemplo, si surgen nuevas técnicas de IA más allá de LLMs en 3 años, evaluar cómo integrarlas).

Criterios de evaluación (mediano plazo): en el año 5 se debe realizar una **evaluación integral:** ¿Se cumplieron las metas del Plan Estratégico a mitad de camino? ¿Qué dice el **Informe del Estado de la Justicia** (publicación periódica en CR) sobre el impacto de la tecnología? ¿Las partes y abogados perciben mejoras en los servicios (posiblemente medido por encuestas de satisfacción)? También se cruzará con los indicadores del Poder Judicial: metas estratégicas de reducción de mora, aumento de productividad por juez, etc., y cuánto contribuyó la IA a ellas. Un criterio central: **ausencia de violaciones de derechos** atribuibles a la IA. Si en 5 años no ha habido un solo caso de que la IA haya causado indefensión o discriminación –y por el contrario hay casos de éxito en acceso a justicia, por ejemplo, personas vulnerables mejor atendidas vía chatbot– entonces el balance será muy positivo.

➤ Visión a Largo Plazo (5-10 años)

A largo plazo, 5-10 años, la visión es de **transformación plena:** la IA integrada de manera madura en la administración de justicia costarricense, con mejoras sostenidas en eficiencia y calidad y con adaptación continua a nuevas circunstancias.

Componentes de la Visión a Largo Plazo:

- **Justicia aumentada, centrada en el ser humano:** en diez años, se espera una administración de justicia donde **cada operador humano cuente con asistentes de IA** que potencien su trabajo. Los jueces dispondrán de análisis jurídicos automatizados al instante, los administradores verán asignaciones optimizadas, y el público tendrá servicios judiciales digitales 24/7 (por ejemplo, portales de resolución de disputas en línea con AI *mediators* para acuerdos en conflictos menores). Todo esto **sin deshumanizar la justicia,**

conservando el contacto humano cuando importa. La visión es que la IA se encargue de la mecánica y el ser humano de la deliberación y la empatía. Costa Rica podría tener incluso un sistema de “**jurisdicción en línea**” para ciertas materias: por ejemplo, trámites monitorios o ejecuciones de pequeña cuantía totalmente electrónicos donde un algoritmo proponga una resolución y un oficial la supervise sumariamente, con posibilidad de apelación a un juez humano solo si una parte disiente,

- **Marco jurídico consolidado y adaptativo:** a 10 años, el marco legal costarricense sobre IA deberá estar **consolidado**. Posiblemente, tras evaluar la experiencia, se habrá decidido si era necesaria alguna reforma constitucional (si en la práctica surgió alguna tensión con la Carta Magna, se podría en este lapso promover una reforma puntual para resolverlo). La legislación y reglamentación secundaria se mantendrán actualizadas –quizá ya incorporando las evoluciones del derecho comparado, como una eventual revisión del EU AI Act o nuevas directrices de la Corte Interamericana de Derechos Humanos en la materia–. La **jurisprudencia nacional** también habrá contribuido: es posible que en 10 años existan sentencias emblemáticas de la Sala Constitucional costarricense delineando principios de “**debido proceso algorítmico**” o similares, las cuales se habrán incorporado al acervo legal. En resumen, un ecosistema normativo estable pero flexible, capaz de absorber las innovaciones que vengan (por ejemplo, si emergen IA cuánticas o formas avanzadas de aprendizaje, que las normas costarricenses puedan darles cabida bajo los mismos principios rectores),
- **Infraestructura y datos de próxima generación:** en el 2035, la infraestructura tecnológica del PJ deberá haberse renovado de acuerdo con la evolución. Esto podría implicar migrar a plataformas aún más avanzadas, manejo de **big data judicial** (quizá integrando datos de otras ramas: registro civil, penitenciario, etc., con los debidos cuidados, para decisiones más informadas), uso de **IA federada** (donde los modelos aprendan de datos sin centralizarlos, para mayor privacidad) y adopción de estándares internacionales de **interoperabilidad**. Costa Rica, al ser parte de organismos como OECD, probablemente siga lineamientos globales de gobierno de datos e IA. La visión es que el Poder Judicial sea tecnológicamente **resiliente y autónomo**: con capacidad interna no solo para usar, sino para crear innovaciones (¿por qué no un laboratorio de IA jurídica que exporte soluciones a otros países?),

- **Usuario empoderado y acceso a la justicia expandido:** uno de los pilares a 10 años es que el ciudadano perciba los beneficios. La IA puede hacer la justicia más accesible: por ejemplo, **traducción automática** en tiempo real en juicios para personas que hablan lenguas indígenas o extranjeras; sistemas de **ayuda para personas sin abogado** que les guíen en la presentación de sus casos; análisis predictivo que permita al Estado identificar necesidades y prevenir conflictos (por ej., detectar patrones en casos de cierta naturaleza e impulsar políticas públicas antes de que lleguen a litigio masivo). En la visión de largo plazo, el Poder Judicial costarricense, apoyado por IA, actúa más proactivamente en la sociedad, no solo resolviendo conflictos sino anticipándolos y contribuyendo a soluciones tempranas (lo cual hilvana con la filosofía de **justicia abierta e innovadora** que ya impulsa el PJ,
- **Adaptación cultural completa:** a esa altura, el **factor generacional** habrá hecho lo suyo: nuevas generaciones de abogados y jueces criados con tecnología verán natural el uso de IA. La cultura organizacional del PJ será la de una institución data-driven (basada en datos) pero con fuerte ética. Las preocupaciones iniciales (sesgos, opacidad) habrán sido en gran medida mitigadas mediante la experiencia y ajustes continuos. La IA será vista simplemente como otra herramienta sofisticada –tal como hoy se ve a los buscadores jurídicos o bases de datos–, y la atención se centrará más en resultados de justicia que en la herramienta en sí.

Hitos Específicos (Largo Plazo):

Para dar seguimiento a esta visión, algunos hitos podrían ser:

- **Año 6-7:** evaluación intermedia global de resultados, ajustes mayores si algo no marcha (por ejemplo, decidir si ampliar el uso de IA a materia penal sustantiva, o si crear un Juzgado 100 % en línea para pequeñas causas a modo experimental),
- **Año 8:** lograr certificaciones o reconocimientos internacionales (por ejemplo, certificación de calidad ISO en sistemas de información judicial con IA, u obtener algún premio de innovación judicial a nivel internacional),
- **Año 10:** publicación de un **informe académico/histórico** que resuma la década de IA en la justicia costarricense, documentando lecciones aprendidas, para beneficio propio y de otros países.

Recursos requeridos: a largo plazo, los recursos deben volverse parte del **presupuesto ordinario** del PJ: la innovación ya no es un proyecto especial, sino un gasto permanente en mantenimiento y actualización de sistemas, capacitación continua del personal (porque habrá rotación y nuevos ingresos que formar), etc. Se podría pensar en un fondo específico o en asegurar que un porcentaje fijo del presupuesto judicial anual se destine a tecnología e IA.

Evaluación de riesgos (largo plazo): en esta etapa, un riesgo es la **obsolescencia**: detenerse y no adoptar una nueva ola tecnológica que pudiera surgir (si la IA actual evoluciona a paradigmas radicalmente nuevos, hay que estar preparados). También, riesgos externos: **ciberamenazas avanzadas**, necesidad de renovar sistemas legados. Pero con 10 años de experiencia, el PJ debería tener la agilidad para responder. Otra cuestión es la **sostenibilidad financiera**: tecnologías pueden abaratarse, pero también surgen costos continuos (licencias, infraestructura). Se debe prever en la planificación financiera de la década. Finalmente, mantener el **equilibrio humano-tecnología**: no caer en complacencia de automatizar de más sin valorar el toque humano; la supervisión ética debe persistir como faro.

Criterios de evaluación (largo plazo): el éxito definitivo se medirá en si la implementación de LLMs e IA **realmente contribuyó a la mejora de la justicia**. Indicadores macro: menor **tasa de congestionamiento judicial**, mayor **confianza de la población en el Poder Judicial** (medida en encuestas nacionales, que esperemos suba al ver una justicia más ágil y accesible), mantenimiento o mejora en la **calidad de las sentencias** (por ejemplo, que no aumente la tasa de revocación en apelación, lo que indicaría que las decisiones siguen siendo sólidas pese a la celeridad). Y, por supuesto, la ausencia de violaciones de derechos fundamentales atribuibles al uso de IA en todo el periodo. Si al cabo de 10 años, Costa Rica puede mostrar que incorporó IA masivamente **sin menoscabar derechos, más bien potenciando la eficacia y equidad**, será un modelo a emular. Esto estaría en línea con los objetivos declarados, tanto a nivel nacional (proteger la dignidad humana en el uso de IA), como internacional (aprovechar IA para mejorar la calidad de la justicia con salvaguardas de transparencia y no discriminación).

En conclusión, esta hoja de ruta integrada en tres ejes –reformas legales, adecuación institucional y cronograma– ofrece un plan **viable y jurídicamente sólido** para que la administración de justicia costarricense adopte modelos de lenguaje de IA. Al fundamentarse en

la Constitución, alinearse con el modelo europeo (Reglamento de IA de la UE, Carta CEPEJ, GDPR) y atender consideraciones prácticas y éticas, la propuesta busca garantizar que la transformación digital judicial ocurra **con seguridad jurídica, responsabilidad algorítmica y fortalecimiento del debido proceso** en Costa Rica. De esta forma, la justicia podrá beneficiarse de las ventajas de la IA (mayor eficiencia, consistencia y accesibilidad) sin renunciar a los valores esenciales que la legitiman ante la ciudadanía: independencia, imparcialidad, transparencia y humanidad en cada decisión.

➤ Apartado Conclusivo

La progresiva incorporación de la inteligencia artificial (IA) en la administración de justicia costarricense, analizada en este capítulo desde las dimensiones constitucional, legal e institucional, pone de relieve la encrucijada que enfrenta el Derecho ante tecnologías disruptivas capaces de reconfigurar sustancialmente la función jurisdiccional. A lo largo de los apartados precedentes, se ha constatado que la IA ofrece promesas de mayor celeridad, eficiencia y racionalidad en la tramitación de los litigios y en la gestión interna de los despachos; sin embargo, también implica retos sin precedentes para principios básicos del Estado de Derecho, tales como la independencia judicial, el debido proceso, la igualdad y la protección de datos personales. Este capítulo ha constatado, de forma pormenorizada, que no basta con un simple entusiasmo tecnocrático: la adopción de sistemas de IA en la judicatura demanda un andamiaje normativo, axiológico e institucional de notable densidad, a fin de salvaguardar la naturaleza irrenunciablemente humana de la potestad de juzgar.

En primer término, el capítulo puso el foco en el “estado actual del marco normativo costarricense”, exponiendo que, pese a las disposiciones ya vigentes sobre protección de datos (Ley N.º 8968) y a algunas iniciativas legislativas en curso (por ejemplo, los Proyectos de Ley N.º 23.771, 23.919 y 24.484), persisten vacíos regulatorios notorios. El texto constitucional, si bien consagra con nitidez la independencia judicial (arts. 9, 154) y la tutela judicial efectiva (art. 41), no acoge de forma explícita la problemática de las decisiones jurisdiccionales automatizadas, lo cual obliga a una hermenéutica que armonice las garantías clásicas con las peculiaridades de la IA. En el artículo 33 (igualdad ante la ley) y en las salvaguardas procesales (art. 39, art. 41), se hallan anclajes suficientes para exigir transparencia y explicabilidad a los algoritmos. No obstante, el

riesgo de “caja negra” algorítmica, la posibilidad de sesgos discriminatorios y la necesidad de mantener una supervisión judicial efectiva subrayan la importancia de legislar de manera más específica.

En lo atinente al **derecho de defensa, el debido proceso y la imparcialidad**, se ha demostrado que el uso de herramientas de IA sin un control humano estricto o sin cauces de impugnación adecuados podría quebrantar garantías fundamentales como la bilateralidad de audiencia y la motivación reforzada de las resoluciones. Los ejemplos comparados –caso Loomis en Wisconsin, o los hallazgos de sesgos raciales en COMPAS– sirven de alerta contra la tentación de delegar en la máquina la “verdad judicial”. El capítulo insistió, pues, en que todo sistema automatizado debe concebirse como un mecanismo de apoyo al juzgador, jamás como sustituto de la prudente deliberación humana.

Parte medular del capítulo examinó la **lección del modelo europeo**, donde el Reglamento de IA (AI Act) en trámite prevé categorizar como “alto riesgo” aquellos usos de IA relacionados con la justicia, imponiendo obligaciones intensas de transparencia, trazabilidad, explicabilidad, control humano y supervisión externa. Desde la **Carta Ética Europea sobre el uso de la IA en los sistemas judiciales** hasta las directrices del Tribunal Europeo de Derechos Humanos (TEDH) sobre no discriminación, se asienta un sólido cuerpo de principios:

1. **Antropocentrismo**: el juez humano sigue siendo el único titular de la potestad de juzgar.
2. **Enfoque de riesgo**: se calibran las exigencias de auditoría y garantía en función del potencial impacto de la IA sobre derechos fundamentales.
3. **Transparencia y explicabilidad**: se pide la apertura comprensible del funcionamiento algorítmico, como vía para no desvirtuar el derecho de defensa y la motivación de las sentencias.
4. **No discriminación**: se exige auditar y corregir posibles sesgos en los datos de entrenamiento.
5. **Supervisión y control institucional**: autoridades nacionales o supranacionales tienen facultades para fiscalizar la fiabilidad de los sistemas, con sanciones disuasorias ante incumplimientos.

Estos elementos concuerdan de manera casi natural con la filosofía garantista de la Constitución costarricense, y, en consecuencia, proporcionan pautas concretas para diseñar o reorientar un marco normativo adaptado a la realidad local. No se trata de importar acríticamente el régimen europeo, sino de inspirarse en sus categorías de “alto riesgo” y en su principio básico de que la IA judicial ha de permanecer subordinada al escrutinio humano y al cumplimiento de derechos fundamentales.

La **segunda sección** del capítulo se centró en los “vacíos y necesidades de reforma”, diagnosticando que, aunque las iniciativas actuales de ley son un paso importante, aún no existe un cuerpo legal articulado que regule las aplicaciones de IA en la judicatura con la profundidad requerida. Se hicieron propuestas concretas:

- La **creación de una ley marco sobre IA** o, en su defecto, la inclusión expresa de la IA judicial como “uso de alto riesgo” en algún texto legal integral, lo que conllevaría obligaciones de auditoría, explicación algorítmica y reserva de ley reforzada,
- **Adaptación de la Ley Orgánica del Poder Judicial** para introducir un capítulo o conjunto de disposiciones que definan la naturaleza servicial de la IA, establezcan el régimen de responsabilidad y obliguen al juez a motivar cómo integra (o descarta) la sugerencia algorítmica,
- **Adecuación de los códigos procesales** para crear cauces de contradicción específicos frente a las salidas automatizadas, a fin de preservar el derecho a la defensa y posibilidad de cuestionar un algoritmo presuntamente sesgado o erróneo.

Este capítulo subrayó la importancia de traducir los principios en instrumentos procesales operativos: por ejemplo, establecer el derecho a solicitar la revisión humana de cualquier dictamen tecnológico que influya en la decisión final, así como la necesidad de asegurar la explicabilidad (“explainability”) para dar pleno cumplimiento a la motivación exigida por la Constitución y la jurisprudencia de la Sala Constitucional.

En el **último tramo**, se abordó cómo estructurar la gobernanza para que la adopción de IA responda a criterios de rigor y legitimidad. Se sugirió:

- **Formalizar una comisión y una unidad especializada en IA judicial**, integrada por magistrados, especialistas en tecnología y en ética, con la misión de fijar lineamientos, evaluar riesgos y autorizar progresivamente la expansión de sistemas de IA en los despachos judiciales,
- **Fortalecer la infraestructura tecnológica** y la formación del personal, evitando depender de algoritmos opacos provistos por terceros. La transparencia algorítmica pasa, en gran medida, por desarrollar o mantener control sobre los modelos de IA, así como por capacitar a jueces y funcionarios para que no se vean superados por la complejidad de la herramienta y sepan fiscalizar su uso,
- **Someter los proyectos de IA a evaluaciones de impacto y auditorías regulares**, incorporando controles *ex ante* y *ex post* que permitan detectar potenciales sesgos o deficiencias.

Con estas medidas, se busca **aligerar la mora judicial** y mejorar la calidad de las resoluciones, sin renunciar al pilar de la independencia judicial ni sacrificar derechos fundamentales. La adopción de una estrategia basada en proyectos piloto, la capacitación continua y la consolidación de un órgano especializado dentro del Poder Judicial, se perfila como una ruta viable para alcanzar una “justicia aumentada” que, en vez de suplantar la discreción jurisdiccional, la complementa y refuerce.

La narrativa vertebradora de este capítulo ha sido la urgencia de conciliar innovación con garantías. Por un lado, **la IA** ofrece soluciones prometedoras para acortar los procesos, reducir rezagos y dotar al sistema de una mayor homogeneidad de criterios. Por el otro, **el orden constitucional** exige preservar la dignidad humana, la autonomía del juzgador y la posibilidad de que cada resolución judicial surja de un razonamiento no meramente algorítmico, sino impregnado de valoración jurídica, motivación y empatía.

La lección de la experiencia europea resulta contundente: la “**human in the loop**” o “supervisión humana efectiva” es la piedra angular para prevenir la sumisión de la judicatura a “oráculos matemáticos” que diluyan la responsabilidad individual y la argumentación. Paralelamente, la introducción de instrumentos de IA debe guiarse por un **enfoque escalonado de riesgo** que imponga mayores cautelas en usos de impacto crítico (p. ej., sugerencias de condenas

penales, clasificación de imputados por probabilidad de reincidencia o decisiones sobre libertades cautelares). Cualquier fallo en estos ámbitos podría suponer un agravio irreparable contra la persona afectada y un golpe a la legitimidad del Poder Judicial.

En esta tesisura, el **ordenamiento costarricense** se encuentra ante la inaplazable labor de reformar las leyes procesales y orgánicas, así como de expedir lineamientos claros que orienten la utilización de la IA en sede judicial. Ello debe complementarse con un cambio cultural profundo, que eduque a los operadores jurídicos en el manejo responsable de la tecnología. Se trata, en definitiva, de construir una confianza pública en la “justicia digital”, donde la ciudadanía perciba que la IA es aliada del juez y no un sustituto, y que su intervención no lesiona la legitimidad ni la imparcialidad de los fallos.

La **conclusión** general que emerge es la viabilidad de una **justicia aumentada** por los recursos digitales, siempre que se adopten salvaguardas robustas y un diseño legal e institucional armonioso. Se pretende no solo agilizar la tramitación, sino, también, consolidar la **seguridad jurídica**, mejorar la consistencia jurisprudencial e impulsar la transparencia. Pero todos estos beneficios dependerán de la calidad con que se legisle, se supervise y se capacite a los operadores.

Si bien la incorporación de IA no es una panacea para la mora ni un sustituto de las reformas de fondo, ofrece un instrumento poderoso que, en manos de un juez formado y respetuoso de los principios constitucionales, potencia las virtudes del sistema. Por el contrario, en ausencia de un “**encuadre jurídico garantista**”, la IA podría introducir nuevos factores de desigualdad e inseguridad. De ahí la insistencia de este capítulo en la necesidad de actuación programática, sosegada y anclada en las garantías fundamentales.

En síntesis, la incorporación de IA en la administración de justicia de Costa Rica, a la luz de la experiencia europea y de las exigencias de nuestro bloque de constitucionalidad, se vislumbra como un cambio inexorable, pero también **oportuno**, para revitalizar el servicio judicial y atender el reclamo ciudadano de eficiencia. Sin embargo, el proceso demanda:

1. **Refuerzo normativo:** con leyes claras que encuadren la IA como tecnología de “alto riesgo” para los derechos, exigiendo control humano y responsabilidad algorítmica.

2. **Institucionalización:** creación de órganos o comisiones que rijan su desarrollo, implanten auditorías y velen por el cumplimiento de los principios éticos.
3. **Capacitación integral:** formar a jueces y funcionarios en la supervisión crítica de los modelos, fomentando una cultura de transparencia y prudencia.

Solo mediante esta convergencia de factores se garantizará un uso de la IA ajustado a los valores fundantes del Estado Social y Democrático de Derecho. **La justicia aumentada**, lejos de substituir la función jurisdiccional, actuará entonces como catalizadora de la modernización judicial, transformando positivamente la relación entre el juez, las partes y el Derecho. Con un marco normativo apropiado y una vigilancia permanente de la ética y los derechos, Costa Rica puede convertirse en referente regional de innovación judicial con rostro humano, asegurando que la tecnología no diluya la sustancia de la tutela judicial efectiva, sino que la consolide en beneficio de la colectividad.

Este subtítulo, por tanto, culmina subrayando la **imperiosa necesidad** de avanzar en una legislación sistemática, en políticas institucionales responsables y en la formación especializada de los operadores, para que los aportes de la IA se traduzcan en una justicia más pronta, eficiente y confiable, sin renunciar a la médula de garantías que legitima cada sentencia y alimenta la esperanza de equidad ante la ley.

3.4.4. Perspectivas del Análisis Económico del Derecho sobre la Implementación Propuesta

La viabilidad y pertinencia de integrar sistemas de inteligencia artificial (IA) en la administración de justicia costarricense no se agotan en su evaluación desde la óptica constitucional y ética. Un análisis comprehensivo demanda, como imperativo metodológico, la incorporación de las herramientas conceptuales provistas por el Análisis Económico del Derecho (AED). Este enfoque, si bien no reduce la función jurisdiccional a una mera ecuación de eficiencia, ofrece una lente pragmática para ponderar la asignación de recursos públicos, anticipar los incentivos generados por la innovación tecnológica y evaluar la racionalidad económica de la

transformación propuesta, siempre en el marco del respeto a los principios fundamentales del Estado de Derecho³²².

La implementación delineada en la hoja de ruta precedente conlleva, inevitablemente, la asunción de costos significativos, tanto directos como indirectos. En primer término, emergen los **costos de adquisición e infraestructura tecnológica**. Estos comprenden no solo el eventual licenciamiento de modelos de IA propietarios, cuyos esquemas de pago recurrentes pueden representar una carga presupuestaria considerable, sino, también, la inversión en hardware especializado. Si bien la adopción de soluciones *open-source* de vanguardia, como el modelo DeepSeek R1, puede mitigar estos gastos al permitir su ejecución local con una inversión inicial relativamente contenida en servidores de alto rendimiento (procesadores EPYC, alta capacidad de RAM DDR5, almacenamiento NVMe), la necesidad de escalar dicha infraestructura para garantizar la concurrencia de usuarios y la redundancia del sistema implica un desembolso inicial relevante. A ello se suman los costos asociados a la adecuación de redes, sistemas de refrigeración y el incremento en el consumo energético.

Secundariamente, deben computarse los **costos de desarrollo e integración**. La adaptación de los modelos de IA, sean estos propietarios o de código abierto, a los sistemas de gestión judicial preexistentes en Costa Rica (como el escritorio virtual o las plataformas de consulta jurisprudencial) exige un esfuerzo considerable en ingeniería de software, desarrollo de interfaces de programación de aplicaciones (APIs) y rigurosas pruebas de interoperabilidad para asegurar una integración fluida y segura.

En tercer lugar, los **costos de mantenimiento y actualización** se perfilan como un factor operativo recurrente. La naturaleza dinámica de la IA impone la necesidad de actualizaciones periódicas de los modelos, reentrenamiento con datos jurídicos actualizados (nueva legislación y jurisprudencia) y soporte técnico continuo, generando así gastos operativos sostenidos en el tiempo.

Cuarto, la estructura de **gobernanza y supervisión** propuesta, incluyendo la Comisión especializada, la Unidad Técnica de IA, la realización sistemática de auditorías algorítmicas y

³²² Robert Cooter y Thomas Ulen, Derecho y Economía, 3^a ed. (Méjico: Fondo de Cultura Económica, 2002); Richard A. Posner, Economic Analysis of Law, 9th ed. (New York: Wolters Kluwer, 2014).

evaluaciones de impacto, si bien indispensable desde una perspectiva garantista, representa una inversión institucional significativa en términos de personal cualificado y recursos dedicados.

Finalmente, los **costos asociados al capital humano y la transición organizacional** no pueden soslayarse. La capacitación intensiva y continua de la judicatura y del personal de apoyo en competencias digitales y éticas, junto con la posible contratación de perfiles técnicos altamente especializados, implica una inversión sustancial. Adicionalmente, la gestión del cambio cultural, la eventual reubicación de personal cuyas tareas se automaticen y la mitigación de riesgos sociales (como la pérdida de confianza pública ante errores algorítmicos) constituyen costos intangibles, pero estratégicamente cruciales.

Frente a este panorama de costos, la implementación de IA judicial promete un abanico de **beneficios potenciales** de gran calado. El más evidente reside en las **ganancias en eficiencia operativa**. La automatización de tareas repetitivas –como la clasificación de escritos, evidenciada en el piloto de Pérez Zeledón, el análisis preliminar de admisibilidad o la generación asistida de borradores en casos estándar– puede acelerar drásticamente la tramitación procesal y contribuir a la reducción de la mora judicial, particularmente en jurisdicciones de litigación masiva como la cobratoria. Asimismo, la optimización algorítmica en la asignación de recursos y la priorización de expedientes podría mejorar la gestión global del sistema.

Más allá de la eficiencia cuantitativa, la IA puede redundar en una **mejora cualitativa de las decisiones judiciales**. El acceso instantáneo y semánticamente enriquecido a vastos corpus de legislación y jurisprudencia, facilitado por LLMs entrenados en el ordenamiento costarricense, puede fortalecer la fundamentación de las resoluciones. En tareas rutinarias, la precisión algorítmica podría minimizar errores humanos, mientras que, en casos estandarizados, la asistencia de IA podría fomentar una mayor coherencia y uniformidad jurisprudencial, robusteciendo así la seguridad jurídica.

Desde la perspectiva del ciudadano, la IA ofrece vías para **mejorar el acceso a la justicia**. Asistentes virtuales o chatbots legales podrían proporcionar orientación básica y facilitar trámites a personas sin representación letrada. Una justicia más célere, además, reduce los costos indirectos asociados al litigio. Finalmente, si se implementa bajo estrictos estándares de explicabilidad, la IA puede incluso **fortalecer la transparencia y la rendición de cuentas**, permitiendo auditorías más

efectivas y haciendo más comprensibles (al menos en parte) los razonamientos asistidos por algoritmos.

La ponderación costo-beneficio desde el AED nos compele a evaluar si estos beneficios justifican la inversión requerida. El dilema sobre el **licenciamiento de software propietario frente a la adopción de modelos *open-source*** adquiere aquí una relevancia económica central. Si bien una solución propietaria podría implicar altos costos recurrentes, la opción de ejecutar modelos abiertos de alto rendimiento en infraestructura local, como se ha descrito, mejora sustancialmente la ecuación económica. El análisis contrafáctico sugiere que los ahorros derivados de la eficiencia operativa podrían compensar la inversión inicial en hardware y personal técnico, especialmente si se compara con los costos sostenidos de mantener estructuras tradicionales con altos niveles de mora. La reubicación y re-capacitación del personal existente, aunque con costos propios, emerge como una alternativa socialmente más viable que una hipotética reducción de plazas.

Crucialmente, el AED no aboga por sacrificar garantías en aras de la eficiencia. La **inversión en un marco regulatorio robusto**, en mecanismos de auditoría, transparencia y supervisión humana –tal como se deriva del modelo europeo y se propone en esta tesis– no debe interpretarse como un lastre económico, sino como una **internalización necesaria de los costos sociales asociados a los riesgos algorítmicos**. Prevenir errores, sesgos discriminatorios, litigios por vulneración de derechos y la erosión de la confianza pública es, en el largo plazo, económicamente más racional que afrontar las consecuencias de una implementación laxa. Las garantías no son un lujo, sino un requisito para la sostenibilidad y legitimidad del sistema.

Asimismo, un marco normativo claro sobre **responsabilidad algorítmica** genera los incentivos correctos para que, tanto proveedoresm como el propio Poder Judicial, actúen con la debida diligencia en el diseño, implementación y supervisión de los sistemas de IA. La consideración de **externalidades positivas** –como la mejora general de la seguridad jurídica que beneficia al clima de negocios, o el incremento del acceso a la justicia para poblaciones vulnerables– refuerza el argumento de que los beneficios sociales de una IA judicial bien implementada trascienden los meros ahorros presupuestarios internos.

En conclusión, desde la perspectiva del Análisis Económico del Derecho, la implementación estratégica de la inteligencia artificial en la administración de justicia costarricense se presenta como una inversión pública potencialmente racional y generadora de valor social neto. Su viabilidad económica, sin embargo, está intrínsecamente condicionada a una gestión prudente que optimice la relación costo-efectividad tecnológica (favoreciendo soluciones abiertas y escalables), priorice la inversión en capital humano (capacitación y especialización), internalice los costos asociados a las salvaguardas ético-jurídicas, y focalice los esfuerzos iniciales en áreas de alto retorno en eficiencia (como la litigación masiva y repetitiva). La eficiencia económica, en este contexto, no puede disociarse de la legitimidad institucional; debe ser un medio para fortalecer una justicia más ágil, consistente y accesible, siempre dentro de los cauces del marco garantista que exige nuestro Estado Constitucional de Derecho.

Conclusiones de la Investigación

En este apartado final se condensan, con el afán de ofrecer una panorámica exhaustiva y un balance crítico, los hallazgos y las reflexiones centrales alcanzados a lo largo de esta investigación. El objetivo primordial que ha guiado la tesis consiste en analizar, críticamente, el marco normativo de la Unión Europea relativo a la implementación de sistemas automatizados de decisión basados en inteligencia artificial (SADIA) en la administración de justicia, para extrapolar lineamientos y mejores prácticas conducentes a la elaboración de una regulación costarricense coherente con los principios y valores del Estado de Derecho. A tal fin, los capítulos previos han efectuado un riguroso escrutinio histórico, técnico, ético y jurídico, examinando, tanto los fundamentos mismos de la inteligencia artificial (Capítulo I), como las disposiciones europeas que pretenden armonizar la innovación tecnológica con la salvaguarda de derechos fundamentales (Capítulo II), para, finalmente, evaluar la situación normativa e institucional costarricense, proyectando recomendaciones viables a corto, mediano y largo plazo (Capítulo III).

La **hipótesis** planteada reza que la incorporación de SADIA en la judicatura costarricense encierra la promesa de agilizar, abaratar y tecnificar múltiples trámites judiciales, muy en particular los de menor complejidad, pero, al mismo tiempo, genera riesgos de sesgos, discriminación y merma del rol esencial del juez como garante de los derechos de las partes. Por ello, la implementación de tales tecnologías debe supeditarse a una regulación específica, inspirada

en los principios constitucionales y en los estándares internacionales, en aras de evitar un uso indiscriminado que vulnere garantías básicas. A la luz de los desarrollos analizados, tanto a nivel europeo como en el contexto costarricense, puede adelantarse que esta hipótesis se **verifica**. En efecto, los datos, las doctrinas y los ejemplos empíricos revisados corroboran el doble filo de estos sistemas: su potencial para elevar la eficiencia judicial y su inherente propensión a replicar prejuicios o vacíos de explicación, si no se les encuadra en un sólido andamiaje normativo y ético.

1. Principales Hallazgos sobre la Evolución Histórica y Técnica de la IA (Relación con el Capítulo I)

El primer capítulo de la tesis expone la trayectoria de la inteligencia artificial desde sus albores hasta las arquitecturas más vanguardistas, como los grandes modelos de lenguaje (LLMs) tipo GPT-4, Claude 3 o Gemini 1.5. Este recorrido demuestra que la IA no constituye un fenómeno repentino, sino el resultado de décadas de investigación teórica, aprendizajes empíricos y avances en el procesamiento de datos a gran escala. Su relevancia para el ámbito judicial deviene patente en la medida en que las redes neuronales profundas y los modelos de lenguaje pueden reducir la sobrecarga de trabajo, sistematizar jurisprudencia, esbozar borradores de resoluciones y automatizar procedimientos repetitivos.

Sin embargo, también se constató que la IA no es inmune a limitaciones: la “opacidad algorítmica” se traduce en la imposibilidad, en ocasiones, de explicar con claridad la lógica subyacente a ciertas decisiones, generando riesgos para la transparencia judicial. Igualmente, si la IA se entrena con datos históricamente sesgados, corre el peligro de perpetuar o exacerbar patrones discriminatorios. Por ende, la adopción de SADIA por parte de un Poder Judicial requiere no solo la pericia técnica para implementar redes neuronales o sistemas de razonamiento simbólico, sino, también, un marco de control institucional que asegure la no vulneración de derechos fundamentales.

Este diagnóstico inicial apuntala el **objetivo específico primero**: “Examinar los instrumentos jurídicos de la Unión Europea que establecen principios éticos y técnicos para la aplicación de la IA en la Justicia”. Aunque tal objetivo se desarrolla en los capítulos subsiguientes, el análisis preliminar del estado del arte tecnológico marca el punto de partida para dimensionar

la complejidad que reviste la regulación de sistemas automatizados. La comprensión de la historia y de las diferentes arquitecturas informáticas constituye la base sobre la cual se erige la reflexión normativa.

2. El Modelo Normativo de la Unión Europea como Referente (Relación con el Capítulo II)

El segundo capítulo procede a un análisis exhaustivo del entramado jurídico europeo, que ha cobrado forma en diversos instrumentos, dentro de los que destacan los siguientes:

1. **Reglamento General de Protección de Datos (RGPD)**: pionero al exigir salvaguardias frente a la toma de decisiones automatizadas con efectos significativos sobre las personas.
2. **Carta Ética Europea sobre el Uso de la IA en la Justicia (CEPEJ)**: establece principios de respeto de los derechos fundamentales, no discriminación, transparencia y control humano.
3. **Reglamento de la Unión Europea sobre IA (AI Act)**: adopta un enfoque basado en niveles de riesgo, catalogando la IA en la justicia como “alto riesgo” y sujetándola a estrictos requisitos de transparencia, supervisión y explicabilidad.

Mediante esta arquitectura normativa, la UE persigue un equilibrio entre la protección de la dignidad humana y la promoción de la innovación: se busca aprovechar la eficiencia de la IA sin sacrificar la legitimidad de la función jurisdiccional. A través de la clasificación por riesgo, el AI Act obliga a los sistemas de IA aplicados a la justicia a someterse a un escrutinio riguroso antes de ser desplegados, asegurando que el juez conserve la autonomía decisoria y que se implementen mecanismos para detectar y corregir sesgos.

Este plexo de normas europeas demuestra que la problemática de la IA judicial excede el ámbito puramente tecnológico, incidiendo en la esencia del Derecho procesal y en la arquitectura institucional de los tribunales. Cabe subrayar que la UE no concibe la IA como un mero suplemento de la justicia, sino como un factor disruptivo que debe encuadrarse cuidadosamente para evitar el debilitamiento de la tutela judicial efectiva. De este modo, el Capítulo II reafirma la **verificación de los objetivos específicos** en relación con la identificación de principios éticos, técnicos y jurídicos a partir de experiencias foráneas (objetivo específico primero y segundo), así

como en lo tocante a la relevancia de la transparencia y el control humano como ejes innegociables (objetivo específico tercero).

3. Diagnóstico de la Situación Costarricense y Propuesta de Lineamientos (Relación con el Capítulo III)

El tercer capítulo centra su atención en el marco costarricense, evidenciando que, si bien la Constitución y el ordenamiento general protegen valores como la igualdad, el debido proceso y la independencia judicial, **no existe todavía** una regulación puntual que establezca las condiciones para el uso legítimo y ético de SADIA en la administración de justicia. Se carece de una ley especializada que contemple las particularidades de la IA, su impacto en los derechos fundamentales y los mecanismos de auditoría preventiva.

El análisis apunta a que, en la actualidad, Costa Rica podría adoptar iniciativas de modernización —por ejemplo, para la clasificación automatizada de escritos judiciales o la sugerencia de borradores de resoluciones simples—, pero el despliegue de sistemas más avanzados (que incidan directamente en la decisión de fondo) se topa con un vacío normativo. Tal vacío se hace más notorio si se confronta con la experiencia europea: mientras el AI Act traza reglas claras de rendición de cuentas y transparencia, en el entorno costarricense no existe un protocolo uniforme de auditoría de algoritmos, ni procedimientos en los que las partes puedan impugnar lo que denominan “alucinaciones” o sesgos del sistema.

Aun con ese panorama, el capítulo concluye que existe suficiente sintonía entre los principios constitucionales de Costa Rica y los postulados europeos de respeto a la dignidad, no discriminación y tutela judicial efectiva, de modo que resultaría factible adaptar los lineamientos del AI Act al medio nacional. Para ello, se sugiere una **hoja de ruta** que comprende:

- 1. Reformas legislativas:** sea la creación de una Ley Marco de Inteligencia Artificial que a la manera europea incorpore una gradación de riesgos en la que se contemple el uso de la IA en la administración de justicia como de “alto riesgo” y, por tanto, se regule el tema de marras tangencialmente a partir de dicha categorización, a falta de una normativa sectorial de IA judicial o la incorporación de capítulos dedicados a la IA en la Ley Orgánica del

Poder Judicial, estableciendo la exigencia de transparencia, el rol indelegable del juez y el derecho de impugnación de las partes frente a errores algorítmicos.

2. **Instalación de una Comisión especializada en la Corte Suprema de Justicia y de una Unidad Técnica Especializada en la Dirección de Tecnologías de la Información del Poder Judicial:** la conformación de una comisión técnica-jurídica de alto nivel, con competencias para autorizar pilotos de IA, velar por la no discriminación y asegurar la supervisión humana.
3. **Capacitación especializada:** invertir en la formación de jueces y funcionarios en IA y ética digital, para que no asuman acríticamente las recomendaciones algorítmicas y retengan su potestad de juzgar con criterio legal y humanista.
4. **Infraestructura tecnológica robusta:** garantizar las condiciones para entrenar y ejecutar modelos bajo altos estándares de ciberseguridad y confidencialidad.

Este abanico de propuestas, además de delinejar el camino normativo-institucional, **ratifica el cumplimiento del objetivo general de la tesis:** brindar un análisis crítico del marco europeo y traducir sus mejores prácticas a la realidad costarricense, ofreciendo así directrices que compatibilicen la potencia transformadora de la IA con los valores básicos que sustentan la función jurisdiccional.

4. Verificación de la Hipótesis de la Investigación

Como se anticipó al comienzo, la hipótesis sostenía que la implementación de SADIA en la administración de justicia costarricense conlleva innegables **aspectos positivos** (optimización de trámites sencillos, reducción de costos, mayor accesibilidad) y **aspectos negativos** (posibles sesgos discriminatorios, disminución de la calidad garantista, sustitución o menoscabo de la figura judicial). Al culminar el estudio, estos extremos se confirman con claridad.

- **Aspectos Positivos:**
 - **Agilidad y eficiencia:** numerosos sistemas basados en IA demuestran un alto rendimiento al procesar grandes volúmenes de datos, facilitando la resolución rápida de litis de cuantía menor o reclamaciones repetitivas,

- **Optimización de recursos:** la automatización de tareas administrativas liberaría tiempo y recursos humanos, que podrían centrarse en casos complejos o en la actividad puramente jurisdiccional,
- **Mayor consistencia en decisiones sencillas:** bien entrenados y supervisados, los algoritmos podrían favorecer una aplicación más uniforme de la legislación en ciertos ámbitos, reduciendo la variabilidad de criterios.

- **Aspectos Negativos y Retos:**

- **Riesgos de sesgos:** el entrenamiento en bases de datos contaminadas por estereotipos históricos o incompletas genera probabilidades de discriminación hacia minorías, reproduciendo desigualdades sociales,
- **Opacidad y falta de explicabilidad:** los mecanismos de redes neuronales profundas —especialmente los modelos cerrados— dificultan la trazabilidad, afectando el derecho de las partes a conocer las razones que fundamentan una decisión,
- **Dilemas éticos y jurídicos:** la función judicial, concebida tradicionalmente como ejercicio reflexivo y valorativo, no se puede descargar meramente en un software. El rol del juez y su potestad decisoria deben preservarse de manera innegociable.

De tal suerte, se confirma que la utilización de SADIA exige un andamiaje normativo riguroso, que oriente cada etapa (desarrollo, entrenamiento, verificación y uso) de la IA, mitigando riesgos y reforzando la legitimidad judicial. Así, la **hipótesis** queda, en términos generales, debidamente **verificada**.

5. Cumplimiento de los Objetivos Planteados

1. Objetivo General:

- *“Analizar críticamente el marco normativo de la Unión Europea relativo a la implementación de SADIA en la justicia, para extraer lineamientos aplicables al sistema costarricense”*,

- El desarrollo de la tesis ha satisfecho con creces este cometido, pues se ha profundizado en la génesis, alcance y filosofía del AI Act y demás instrumentos europeos, a la par de ponerlos en cotejo con la realidad costarricense. Además, se formulan recomendaciones viables para concretar dicha adaptación.

2. Objetivos Específicos:

- a) *“Examinar los instrumentos jurídicos de la Unión Europea que establecen principios éticos y técnicos para la aplicación de la IA en la Justicia”.*
 - Se ha realizado un escrutinio abarcador del RGPD, la Carta Ética de la CEPEJ, las directrices para la presentación electrónica de documentos judiciales y, sobre todo, del AI Act. Se delimitan los criterios de clasificación por riesgo, las obligaciones de transparencia y la supervisión humana, quedando por ende cumplido el objetivo.
- b) *“Identificar los principales beneficios y riesgos asociados a la implementación de sistemas automatizados de decisión en la resolución de conflictos judiciales, con base en experiencias documentadas en los Estados Miembros de la UE”.*
 - El recorrido por la normatividad y la jurisprudencia europea, así como por ejemplos concretos (casos de automatización parcial en procesos de menor cuantía en Estonia, la identificación biométrica discutida en Francia, los proyectos piloto en España, entre otros), ha permitido extraer los aspectos positivos y negativos de la IA judicial, satisfaciendo asimismo este segundo objetivo.
- c) *“Formular una propuesta de lineamientos jurídicos y buenas prácticas para encauzar responsablemente un eventual proceso de incorporación de estas tecnologías en el sistema judicial costarricense”.*
 - El Capítulo III, sirviéndose de la metodología comparada y el análisis de vacíos legales, sugiere una batería de reformas legislativas, estructurales e institucionales. Estas recomendaciones guardan consonancia con los principios rectores defendidos por la UE, completando así el tercer objetivo.

En consecuencia, puede afirmarse que la investigación no solo alcanza los objetivos planteados, sino que, además, provee una reflexión comprehensiva que combina la teoría, la praxis comparada y la propuesta regulatoria contextualizada para Costa Rica.

6. Reflexiones Finales y Proyección de Futuro

Al margen de la constatación de los beneficios y retos que supone el uso de la IA en la justicia, la gran conclusión que emana de esta tesis es que cualquier adopción de SADIA debe estar guiada por el **imperativo de salvaguardar** la independencia e imparcialidad judicial, la transparencia del proceso y el derecho a un recurso efectivo. No basta con la pericia tecnológica; se requiere un **marco ético-normativo** robusto y una adecuada dotación institucional para que la IA se convierta en un recurso que potencie la función jurisdiccional y no en un factor que la opague o la desnaturalice.

Por ende, la apuesta más sensata es la de la “**justicia aumentada**”, en la que la IA se conciba como un auxiliar que optimiza tareas repetitivas, reduce congestiones y asiste en el análisis de grandes volúmenes de datos, sin desplazar la conciencia crítica del juez ni la facultad de las partes de contradecir los medios de prueba o las sugerencias automatizadas. Resulta igualmente imprescindible que, en casos de menor complejidad, se valore la posible automatización parcial, siempre que un magistrado u órgano revisor retenga la última palabra y que existan vías procesales expeditas para impugnar cualquier disfunción algorítmica.

Dado el ritmo vertiginoso de la evolución tecnológica, es plausible que, en el mediano plazo, los grandes modelos de lenguaje y otros sistemas de IA logren niveles de precisión y coherencia argumentativa cada vez mayores. Con todo, ello no disipa la necesidad de **normas claras** que establezcan quién asume responsabilidad legal por eventuales daños causados por la máquina, de qué modo se salvaguarda la privacidad de los datos judiciales y cómo se garantiza la transparencia frente a las partes. La experiencia europea exhibe un camino de regulación *ex ante* que, lejos de constituir un freno a la innovación, la encauza bajo un criterio de confiabilidad y legitimidad.

Así, esta investigación contribuye a perfilar una ruta para Costa Rica, según la cual, el país podría transitar, con prudencia y rigor, de la fase exploratoria a la adopción formal de SADIA.

Dicho tránsito, de concretarse con el respaldo legislativo y la coordinación de altos órganos judiciales, podría otorgarle al Poder Judicial un grado de modernización acorde con la dinámica del siglo XXI, sin, por ello, desatender el peso de los principios constitucionales que constituyen el basamento de la justicia costarricense.

En suma, y a modo de **corolario**:

- Se constata la urgente necesidad de regular la IA antes de que su implementación crezca de manera desordenada,
- El **modelo europeo**, con su énfasis en el “riesgo alto” de la IA judicial y la obligación de control humano, constituye un **referente sólido** para la elaboración de una normativa costarricense,
- La **hipótesis** se confirma, al acreditarse tanto los beneficios potenciales de la IA como sus riesgos en la administración de justicia,
- Los **objetivos** planteados —tanto el general como los específicos— han sido cumplidos mediante la revisión crítica de fuentes europeas, el análisis comparativo y la formulación de lineamientos concretos para Costa Rica.

Por ende, este trabajo concluye enfatizando que la convergencia entre Derecho e IA no puede reducirse a una visión meramente instrumental de la tecnología. Antes bien, su éxito o fracaso reside en la compatibilidad de los sistemas automatizados con el entramado de valores que distinguen a la función judicial como bastión de la democracia, la seguridad jurídica y la dignidad humana. En la medida en que se conciban y apliquen **políticas públicas y reformas legislativas** inspiradas en la transparencia, la explicabilidad, la responsabilidad y la supervisión judicial, será posible aprovechar cabalmente el potencial de la IA para brindar una justicia más eficiente y accesible, sin comprometer los principios nucleares que legitiman el acto de juzgar. Sin embargo, el éxito de la implementación de IA en la justicia costarricense dependerá no solo de la adopción de principios garantistas inspirados en el modelo europeo, sino, también, de la configuración de un **marco normativo flexible y adaptable**, que priorice la regulación principista en la ley y delegue los detalles técnicos a normativa secundaria, así como del establecimiento de **órganos de gobernanza institucional ágiles y técnicamente competentes**, capaces de responder con celeridad a los desafíos de esta tecnología en constante evolución.

Bibliografía

Almeida, Ines. "Is GPT-4 a Mixture of Experts Model? Exploring MoE Architectures for Language Models". AI Insights (blog), 17 de agosto de 2023. <https://wwwnownextlaterai.com/insights/gpt-4-moe>.

Alonso García, Ricardo. "El Soft Law Comunitario". *Revista de Administración Pública*, no. 154 (2001): 63-94. <https://www.cepc.gob.es/sites/default/files/2021-12/243532001154063.pdf>

Angwin, Julia, Jeff Larson, Surya Mattu, y Lauren Kirchner. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks". ProPublica, 23 de mayo de 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Anthropic. Claude 2: The Next Leap Forward. 2023. <https://www.anthropic.com/news/clause-2>

Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Reporte técnico, marzo de 2024. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

Anthropic. "Introducing the Next Generation of Claude". 4 de marzo de 2024. <https://www.anthropic.com/news/clause-3-family>

Anthropic. "Long context prompting for Claude 2.1". 6 de diciembre 2023. <https://www.anthropic.com/news/clause-2-1-prompting>

Asamblea Legislativa de la República de Costa Rica. *Proyecto de Ley – Ley de Ciberseguridad de Costa Rica, Expediente N.º 23.292*. Presentado por José Joaquín Hernández Rojas y varios señores y señoras diputados. Recuperado de: https://cicr.com/wp-content/uploads/2022/10/Exp_23292.pdf

Asamblea Legislativa de la República de Costa Rica. *Proyecto de Ley N.º 24.484: Ley para la Implementación de Sistemas de Inteligencia Artificial (IA) (2025)*.

Asamblea Legislativa de la República de Costa Rica. *Proyecto de Ley N.º 23.919: Ley para la Promoción Responsable de la IA en Costa Rica (2024)*.

Asamblea Legislativa de la República de Costa Rica. *Proyecto de Ley N.º 23.771, Ley de Regulación de la Inteligencia Artificial en Costa Rica (2023)*.

Banco Interamericano de Desarrollo (BID). *Costa Rica promoverá el uso responsable de la inteligencia artificial con apoyo del BID*. Comunicado de prensa, San José, 29 de septiembre de 2021. Recuperado de: <https://www.iadb.org/es/noticias/costa-rica-promovera-el-uso-responsable-de-la-inteligencia-artificial-con-apoyo-del-bid#:~:text=San%20Jos%C3%A9%2C%2029%20de%20septiembre,Instituto%20Tecnol%C3%B3gico%20de%20Costa>

Barrio Andrés, Moisés, dir. *El Reglamento Europeo de Inteligencia Artificial*. Valencia: Tirant lo Blanch, 2024.

Barona Vilar, Silvia. "Dataización de la justicia (algoritmos, inteligencia artificial y justicia, ¿el comienzo de una gran amistad?)". *Revista Boliviana de Derecho*, no. 36 (julio 2023): 14–45. ISSN: 2070-8157. Recuperado de: <https://dialnet.unirioja.es/descarga/articulo/9043836.pdf>

Bedayn, Jesse. "Attempts to Regulate AI's Hidden Hand in Americans' Lives Flounder in US Statehouses". Associated Press, 23 de mayo de 2024. https://apnews.com/article/artificial-intelligence-bias-discrimination-regulation-ai-ff1d0860663723079aac3666b38f2320?utm_source=chatgpt.com.

Boletín Oficial de las Cortes Generales. Congreso de los Diputados. XIV Legislatura. Serie A: Proyectos de Ley, 8 de junio de 2023. Informe de la Ponencia Núm. 97-4, 121/000097

Proyecto de Ley de medidas de eficiencia procesal del servicio público de Justicia.
https://www.congreso.es/public_oficiales/L14/CONG/BOCG/A/BOCG-14-A-97-4.PDF.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, et al. "On the Opportunities and Risks of Foundation Models". arXiv:2108.07258v3 [cs.LG], 12 de julio de 2022. 2108.07258.pdf (arxiv.org).

Borges Blázquez, Raquel. "El sesgo de la máquina en la toma de decisiones en el proceso penal". IUS ET SCIENTIA 6, no. 2 (2020): 54-71. Universidad de Sevilla, España.
<https://revistascientificas.us.es/index.php/ies/article/view/14328/12770>

Brachman, Ronald J., y Hector J. Levesque, con una contribución de Maurice Pagnucco. Knowledge Representation and Reasoning. San Francisco, CA: Morgan Kaufmann Publishers,

2004.

<https://www.cin.ufpe.br/~mtcfa/files/in1122/Knowledge%20Representation%20and%20Reasoning.pdf>

Bradford, Anu. *The Brussels Effect: How the European Union Rules the World*. Nueva York: Oxford University Press, 2020. <https://dokumen.pub/qdownload/the-brussels-effect-how-the-european-union-rules-the-world-9780190088583-2019031328-2019031329-9780190088606-9780190088590-9780190088613.html>

Branwen, Gwern. "GPT-3 Creative Fiction". 2020. <https://www.gwern.net/GPT-3>

Brenes Mora, Samantha. "Micitt lanza primera política pública sobre uso y desarrollo de Inteligencia Artificial". Delfino.cr, 24 de octubre de 2024, 3:54 p. m. Recuperado de: <https://delfino.cr/2024/10/micitt-lanza-primera-politica-publica-sobre-uso-y-desarrollo-de-inteligencia-artificial>

Brown, Tom B., et al. "Language models are few-shot learners". arXiv preprint arXiv:2005.14165, 2020. <https://arxiv.org/abs/2005.14165>

Buchanan, Bruce G. "A (very) brief history of artificial intelligence". AI Magazine 26, no. 4 (2005): 53-60. <https://doi.org/10.1609/aimag.v26i4.1848>

Cai, Culhong, y Jiahui Yin. "Cultural and Ethical Foundations of AI Governance Divergence: A Comparative Analysis of China and the West". Política Internacional VII, no. 1 (enero-marzo, 2025): 215-233. https://cas.fudan.edu.cn/info/1212/21153.htm?utm_source=chatgpt.com

Campbell, A. N., V. F. Hollister, R. O. Duda, y P. E. Hart. "Recognition of a Hidden Mineral Deposit by an Artificial Intelligence Program". Science 217, no. 4563 (1982): 927-929. https://www.jstor.org/stable/1689346?oauth_data=eyJlbWFpbCI6ImtzYW5jaGV6emFtb3JhQGdtYWlsLmNvbSIzImluc3RpdHV0aW9uSWRzIjpBXSwicHJvdmlkZXIiOiJnb29nbGUifQ

Carrigan, Matthew (@carrigmat). "Complete hardware + software setup for running Deepseek-R1 locally. The actual model, no distillations, and Q8 quantization for full quality. Total cost, \$6,000. All download and part links below:". Hilo en X, 28 de enero de 2025, 8:17 a. m. Recuperado de: <https://x.com/carrigmat/status/1884244369907278106>

Chan, Terry E., Elizabeth M. Beh, Vishal Dalal, Sandeep Ganesh, Jing Li, y Andelyn Russell. "Technology and Geopolitics: What If The Semiconductor Industry Bifurcates?". S&P Global Ratings, 14 de noviembre de 2022. <https://www.spglobal.com/ratings/en/research/articles/221114-technology-and-geopolitics-what-if-the-semiconductor-industry-bifurcates-12557030>

Chen, Benjamin Minhao, y Zhiyu Li. "How Will Technology Change the Face of Chinese Justice?". Columbia Journal of Asian Law 34, no. 1 (2020): 1-21. <https://www.google.co.cr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjYpaXcjaKGAXERDABHQ4wDgsQFnoECA4QAQ&url=https%3A%2F%2Fjournals.library.columbia.edu%2Findex.php%2Fcjal%2Farticle%2Fdownload%2F7484%2F3923%2F14211&usg=AOvVaw3C4YIaB6W8Hd4WNo5BrfsZ&opi=89978449>

Comisión Europea. Evaluación de impacto: Acompañando la Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas sobre inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión. SWD(2021) 84 final. Bruselas, 21 de abril de 2021. <https://digital-strategy.ec.europa.eu/en/library/impact-assessment-regulation-artificial-intelligence>

Comisión Europea. Libro Blanco sobre la Gobernanza Europea. EUR-Lex: Acceso al Derecho de la Unión Europea. Última modificación el 21 de febrero de 2008. <https://eur-lex.europa.eu/ES/legal-content/summary/white-paper-on-governance.html>

Comisión Europea. Libro Blanco sobre la inteligencia artificial: un enfoque europeo orientado a la excelencia y la confianza. COM(2020) 65 final, Bruselas, 19 de febrero de 2020. <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52020DC0065>

Comisión Europea. "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." COM(2021) 206 final, 21 de abril de 2021. <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52021PC0206&from=EN>

Comisión Europea. Revisión de 2021 del plan coordinado sobre la inteligencia artificial: Comunicación de la Comisión al Parlamento Europeo, al Consejo Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. COM(2021) 205 final, Bruselas, 21 de abril de 2021. <https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review>

Comisión Europea para la Democracia a través del Derecho (Comisión de Venecia). Informe preliminar sobre la independencia del sistema judicial: Parte I: La independencia de los jueces, Estudio No. 494/2008, Estrasburgo, 5 de marzo de 2010, CDL(2010)006. Recuperado de: [https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL\(2010\)006-e](https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL(2010)006-e)

Comité de Derechos Humanos. Observación general núm. 32, artículo 14: El derecho a un juicio imparcial y a la igualdad ante los tribunales y cortes de justicia, 2007. Recuperado de: <https://www.refworld.org/es/leg/coment/ccpr/2007/es/52583>

Consejo de Europa. "Convenio Europeo de Derechos Humanos", 1950. https://www.echr.coe.int/documents/convention_spa.pdf.

Consejo de Europa. "European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment." Adoptada en la 31^a reunión plenaria de la CEPEJ, Estrasburgo, 3-4 de diciembre de 2018. <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>

Corte Interamericana de Derechos Humanos. Caso Apitz Barbera y otros ("Corte Primera de lo Contencioso Administrativo") vs. Venezuela, Sentencia de 5 de agosto de 2008. Recuperado de: https://www.corteidh.or.cr/docs/casos/articulos/seriec_182_esp.pdf

Corvalán, Juan Gustavo. Prometea: Inteligencia Artificial para Transformar Organizaciones Públicas. Buenos Aires: Editorial Astrea, 2019. Conferencia durante la Asamblea Ordinaria del Consejo Permanente de la Organización de los Estados Americanos, 22 de agosto de 2018, Washington D.C. <http://scm.oas.org/pdfs/2018/CP-PRES-CORV.pdf>

Davis, Randall, y Jonathan J. King. "The Origin of Rule-Based Systems in AI". En Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, editado por B . G. Buchanan y E. H. Shortliffe, 20-52. Addison-Wesley, 1984. <https://www.shortliffe.net/Buchanan-Shortliffe-1984/MYCIN%20Book.htm>

De Asis Pulido, Miguel. "La justicia predictiva: tres posibles usos en la práctica jurídica". En Inteligencia Artificial y Filosofía del Derecho, dirigido por Fernando H. Llano Alonso, coordinado por Joaquín Garrido Martín y Ramón Valdivia Jiménez. Murcia: Ediciones Laborum, 2022.

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv preprint arXiv:2501.12948v1, 22 de enero de 2025.
<https://arxiv.org/abs/2501.12948>

Dewan, Shaila. "Judges Replacing Conjecture With Formula for Bail". The New York Times, 16 de junio de 2015. <https://www.nytimes.com/2015/06/27/us/turning-the-granting-of-bail-into-a-science.html>

"Dot CSV". "¿Qué es una Red Neuronal? Parte 1: La Neurona | DotCSV". YouTube Video, 2:00. 19 de marzo de 2018. <https://www.youtube.com/watch?v=MRIv2IwFTPg&list=PL-Ogd76BhmcB9OjPucsnc2-piEE96jJDQ>

"Dot CSV". "¿Qué es una Red Neuronal? Parte 2: La Red | DotCSV". YouTube Video, 1:53. Publicado el 26 de marzo de 2018.
<https://www.youtube.com/watch?v=uwbHOpp9xkc&list=PL-Ogd76BhmcB9OjPucsnc2-piEE96jJDQ&index=2>

"Dot CSV". "¿Qué es una Red Neuronal? Parte 3: Backpropagation | DotCSV". YouTube Video. Publicado el 3 de octubre de 2018.
https://www.youtube.com/watch?v=eNIqz_noix8&list=PL-Ogd76BhmcB9OjPucsnc2-piEE96jJDQ&index=4

Dreyfus, Hubert L. What Computers Can't Do: The Limits of Artificial Intelligence. New York: Harper & Row, 1972.
https://monoskop.org/images/c/ce/Dreyfus_Hubert_What_Computers_Cant_Do_A_Critique_of_Artificial_Reason.pdf

Dworkin, Ronald. Taking Rights Seriously. Traducido por Marta Guastavino. Londres: Gerald Duckworth & Co. Ltd., 1984. 2^a edición, diciembre 1989, ISBN 84-344-1508-9.
<https://img.lpderecho.pe/wp-content/uploads/2021/09/Descargue-en-PDF-Los-derechos-en-serio-de-Ronald-Dworkin-LP.pdf>

East China University of Political Science and Law, Digital Rule of Law Institute. "Artificial Intelligence Law of the People's Republic of China (Draft for Suggestions from Scholars) [中华人民共和国人工智能法 (学者建议稿)]." Traducido por Etcetera Language Group, Inc., editado por Ben Murphy. CSET Translation Manager, 2 de mayo de 2024. https://cset.georgetown.edu/publication/china-ai-law-draft/?utm_source=chatgpt.com

European Commission for the Efficiency of Justice (CEPEJ). European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment. Estrasburgo: Council of Europe, 2018. <https://www.europarl.europa.eu/cmsdata/196205/COUNCIL%20OF%20EUROPE%20-%20European%20Ethical%20Charter%20on%20the%20use%20of%20AI%20in%20judicial%20systems.pdf>

Feigenbaum, E. A. The Art of Artificial Intelligence: I. Themes and Case Studies of Knowledge Engineering. Stanford, CA: Departamento de Ciencias de la Computación, Universidad de Stanford, agosto de 1977. Memo HPP-77-25, Núm. de reporte STAN-CS-77-621. <https://stacks.stanford.edu/file/druid:bg342cm2034/bg342cm2034.pdf>

Floridi, Luciano, y Massimo Chiriatti. "GPT-3: Its Nature, Scope, Limits, and Consequences". Minds and Machines 30, no. 681 (2020): 681-694. <https://doi.org/10.1007/s11023-020-09548-1>

Fridman, Lex. "Dario Amodei: Anthropic CEO on Claude, AGI & the Future of AI & Humanity". Episodio 452 de Lex Fridman Podcast, 54:49. Video de YouTube, publicado el 11 de noviembre de 2024. <https://www.youtube.com/watch?v=ugvHCXCOmm4&t=3581s>

García, Regina. "¿La inteligencia artificial tiene sesgos?". Instituto Mexicano para la Competitividad A.C., 8 de febrero de 2023. <https://imco.org.mx/la-inteligencia-artificial-tiene-sesgos/>

Gemini Team. "Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context". Google DeepMind, 2024. https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf

Graham Greenleaf. "Global Data Privacy Laws 2021: Despite COVID Delays, 145 Laws Show GDPR Dominance". Privacy Laws & Business International Report, no. 169 (2021): 1-5. Recuperado de: ssrn_id3933588_code722134.pdf (elsevier-ssrn-document-store-prod.s3.amazonaws.com).

Gravett, Willem H. "Judicial Decision-Making in the Age of Artificial Intelligence". En Multidisciplinary Perspectives on Artificial Intelligence and the Law, editado por Henrique Sousa Antunes, Pedro Miguel Freitas, Arlindo L. Oliveira, Clara Martins Pereira, Elsa Vaz de Sequeira, y Luís Barreto Xavier, 277-290. Cham, Suiza: Springer Nature Switzerland AG, 2024. <https://doi.org/10.1007/978-3-031-41264-6>

Gruber, Thomas R. "A Translation Approach to Portable Ontology Specifications". Knowledge Acquisition 5, no. 2 (1993): 199-220. <https://doi.org/10.1006/knac.1993.1008>

Harmon, Paul, y David King. Expert Systems: Artificial Intelligence in Business. Wiley, 1985. <https://momot.rs/d3/y/1710919988/107/u/annas-archive-ia-2023-06-lcpdf/e/expertsystemsart00harm.pdf~tl2NkbHLmnQ9TkmimMJsIw/Expert%20systems%3A%20artificial%20intelligence%20in%20business%20-%20Harmon%2C%20Paul%2C%201942-%3B%20King%2C%20David%2C%201949-%20-%204.print.%2C%20New%20York%2C%201985%20-->

[%20New%20York%3A%20J.%20Wiley%20--%209780471808244%20--%200500ae46cb7987e7b743c0ecbc17cb2b%20--%20Anna%E2%80%99s%20Archive.pdf](#)

Hart, H.L.A. The Concept of Law. 2^a ed. Oxford: Oxford University Press, 1961. [https://annas-archive.org/md5/fff38067a2ea7d435f3a14f0e2d27d88](#)

Haugeland, J. Artificial Intelligence: The Very Idea. MIT Press, 1985.
[https://terrorgum.com/tfox/books/artificialintelligence_theveryidea.pdf](#)

Heikkilä, Melissa. "How's AI Self-Regulation Going? One Year on from the White House's Voluntary Commitments on AI". MIT Technology Review, 23 de julio de 2024.
[https://www.technologyreview.com/2024/07/23/1095218/hows-ai-self-regulation-going/?utm_source=chatgpt.com](#)

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, y Hany Hassan Awadalla. "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation". Microsoft, febrero de 2023. DOI: 10.48550/arXiv.2302.09210,
[https://www.researchgate.net/publication/368664574_How_Good_Are_GPT_Models_at_Machine_Translation_A_Comprehensive_Evaluation/fulltext/63f4374f57495059452fbe19/How-Good-Are-GPT-Models-at-Machine-Translation-A-Comprehensive-Evaluation.pdf?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19](#)

Hildebrandt, Mireille. "Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics". University of Toronto Law Journal 68, no. S1 (2018): 12-35. [https://doi.org/10.3138/utlj.2017-0044](#)

Hill, Kashmir. "The Secretive Company That Might End Privacy as We Know It". The New York Times, 18 de enero de 2020. <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>

Hutchins, J. "The Georgetown-IBM experiment demonstrated in January 1954". En Machine Translation: From Real Users to Research: 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, September 28 – October 2, 2004, editado por Robert E. Frederking y Kathryn B. Taylor, 102-114. Berlin: Springer Verlag, 2004. <https://aclanthology.org/www.mt-archive.info/00/AMTA-2004-Hutchins.pdf>

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, et al. "Predicción de estructuras de proteínas altamente precisa con AlphaFold". Nature 596, n.º 7873, 26 de agosto de 2021: 583-589. <https://www.nature.com/articles/s41586-021-03819-2>

Katz, Daniel Martin. "Quantitative Legal Prediction—or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry." Emory Law Journal 62, no. 4 (2013): 909–66. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2187752

KPMG. How the EU AI Act Affects US-Based Companies: A Guide for CISOs and Other Business Leaders. KPMG LLP, marzo de 2024. <https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2024/decoding-eu-ai-act.pdf>

Krizhevsky, Alex, Ilya Sutskever, y Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". Advances in Neural Information Processing Systems 25 (2012): 1097-1105. <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>

Kurzweil, Ray. *The Age of Spiritual Machines*. Barcelona: Editorial Planeta, S.A., 1999; reimpresión exclusiva para México, México, D.F.: Editorial Planeta Mexicana, S.A. de C.V., 2000.

LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, y Lawrence D. Jackel. "Backpropagation applied to handwritten zip code recognition". *Neural Computation* 1, no. 4 (1989): 541-551. <https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf>

LeCun, Yann, Léon Bottou, Yoshua Bengio, y Patrick Haffner. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE* 86, no. 11 (1998): 2278-2324. https://axon.cs.byu.edu/~martinez/classes/678/Papers/Convolution_nets.pdf

LeCun, Yann, Yoshua Bengio, y Geoffrey Hinton. "Deep learning". *Nature* 521, no. 7553 (2015): 436-444. <https://www.nature.com/articles/nature14539>

Lex Machina. "Predict the Behavior of Courts, Judges, Lawyers and Parties with Legal Analytics". Lex Machina. Accedido el 22 de mayo de 2024. <https://www.lexmachina.com>

Lighthill, J. "Artificial Intelligence: A General Survey". En *Artificial Intelligence: A Paper Symposium*, 1-29. Science Research Council, 1973. https://rodsmithe.nz/wp-content/uploads/Lighthill_1973_Report.pdf

Lord Chancellor y Secretario de Estado de Justicia. "Transformando nuestro sistema de justicia: estrategia digital asistida, convicción automática en línea y pena estándar estatutaria, y composición de paneles en tribunales. Respuesta del gobierno". Presentado al Parlamento por mandato de Su Majestad, febrero de 2017. Ministerio de Justicia Británico. transforming-our-justice-system-government-response.pdf (publishing.service.gov.uk).

MacCormick, Neil. *Legal Reasoning and Legal Theory*. Oxford: Oxford University Press, 1978.

Madiega, Tambiama, y Anne Louise Van De Pol. Artificial Intelligence Act and Regulatory Sandboxes: Summary. EPRS | European Parliamentary Research Service, Members' Research Service, Informe PE 733.544, junio de 2022.
[https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI\(2022\)733544_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf)

Marcus, Gary, y Ernest Davis. "GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About". MIT Technology Review, 22 de agosto de 2020.
<https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>

Marín Mena, Andrea. "Poder Judicial avanza en la innovación de sus servicios a través de su fortalecimiento tecnológico". Poder Judicial de Costa Rica, 2022. Recuperado de:
<https://cij.poder-judicial.go.cr/index.php/services/noticias/item/50-poder-judicial-avanza-en-la-innovacion-de-sus-servicios-a-traves-de-su-fortalecimiento-tecnologico#:~:text=que%20atiende%20el%20Poder%20Judicial>

May Grosser, Sebastian. "Gobierno presentó Estrategia Nacional de Ciberseguridad 2023 2027". Delfino CR, 13 de noviembre de 2023. Recuperado de: <https://delfino.cr/2023/11/gobierno-presento-estrategia-nacional-de-ciberseguridad-2023-2027>

McCarthy, John. "Recursive Functions of Symbolic Expressions and Their Computation by Machine, Part I". Communications of the ACM 3, no. 4 (1960): 184-195.
<https://doi.org/10.1145/367177.367199>

McCulloch, Warren S., y Walter Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity". Bulletin of Mathematical Biology 52, nos. 1/2 (1990): 99-115.
<https://www.cs.cmu.edu/~epxing/Class/10715/reading/McCulloch.and.Pitts.pdf>

McDermott, D., y J. Doyle. "Non-Monotonic Logic I". Artificial Intelligence 13, no. 1-2 (1980): 41-72. <https://www.sciencedirect.com/science/article/pii/0004370280900120>

McDermott, J. "R1: A Rule-Based Configurer of Computer Systems". Artificial Intelligence 19, no. 1 (1980): 39-88. main.pdf (sciencedirectassets.com)

Meta. "Presentamos Meta Llama 3: modelo de lenguaje a gran escala más potente hasta la fecha". Publicado el 18 de abril de 2024. <https://www.meta.com/news/meta-llama-3>

Meta Llama. "Responsible Use Guide: Your Resource for Building Responsibly". <https://www.meta.com/responsible-use-guide>

Milgram, Anne. "Why Smart Statistics Are the Key to Fighting Crime". TED@BCG San Francisco. Filmado en octubre de 2013. TED Video. https://www.ted.com/talks/anne_milgram_why_smart_statistics_are_the_key_to_fighting_crime

Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones (MICITT). Estrategia Nacional de Inteligencia Artificial de Costa Rica. San José, C.R.: MICITT, 2024. ISBN 978-9968-732-94-9. Recuperado de: <https://www.micitt.go.cr/sites/default/files/2024-10/Estrategia%20Nacional%20de%20Inteligencia%20Artificial%20de%20Costa%20Rica%20ESP.pdf>

Minsky, Marvin. A Framework for Representing Knowledge. MIT-AI Laboratory Memo 306, junio de 1974. Reimpreso en The Psychology of Computer Vision, ed. P. Winston. McGraw-Hill, 1975. <https://courses.media.mit.edu/2004spring/mas966/Minsky%201974%20Framework%20for%20knowledge.pdf>

Miranda Bonilla, H. "El derecho de acceso a internet en la jurisprudencia de la sala constitucional de Costa Rica". Revista Jurídica Mario Alario D'Filippo 13, no. 25 (2021): 5–18.
<https://doi.org/10.32997/2256-2796-vol.13-num.25-2021-3610>

Mistral AI Team. "Mixtral of Experts: A High Quality Sparse Mixture-of-Experts". 11 de diciembre de 2023. <https://mistral.ai/news/mixtral-of-experts/>

Moor, James. "La conferencia de inteligencia artificial del Dartmouth College: los próximos cincuenta años". AI Magazine 27, núm. 4 (2006): 87-91.
<https://onlinelibrary.wiley.com/doi/pdf/10.1609/aimag.v27i4.1911>

Moore, Gordon E. "Cramming More Components onto Integrated Circuits". Electronics 38, no. 8 (19 de abril de 1965). <https://www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf>

Murillo, Álvaro. "Gobierno lanza estrategia nacional sobre Inteligencia Artificial". Semanario Universidad, 24 de octubre de 2024. Recuperado de:
<https://semanariouniversidad.com/pais/gobierno-lanza-estrategia-nacional-sobre-inteligencia-artificial/#:~:text=La%20ENIA%20procura%20pautas%20t%C3%A9cnicas,materia%3A%20Chile%2C%20Brasil%20y%20Uruguay>

Naciones Unidas. Principios básicos relativos a la independencia de la judicatura, 1985. Recuperado de: <https://www.ohchr.org/es/instruments-mechanisms/instruments/basic-principles-independence-judiciary>

National Museum of American History. "Altair 8800 Microcomputer". Smithsonian.
https://americanhistory.si.edu/collections/nmah_334396

Newell, Allen, y Herbert A. Simon. "GPS, a program that simulates human thought". En Lernende Automaten, ed. H. Billing, 109-124. Oldenbourg, 1961.
https://iiif.library.cmu.edu/file/Simon_box00064_fld04907_bdl0001_doc0001/Simon_box00064_fld04907_bdl0001_doc0001.pdf

Newell, Allen, J.C. Shaw, y Herbert A. Simon. "Elements of a theory of human problem solving". Psychological Review 65, no. 3 (1958): 151-166.
https://iiif.library.cmu.edu/file/Simon_box00064_fld04878_bdl0001_doc0001/Simon_box00064_fld04878_bdl0001_doc0001.pdf

Nissan, Ephraim. "Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement". AI & Society 32 (2017): 441-464.
<https://link.springer.com/article/10.1007/s00146-015-0596-5>

OpenAI. "GPT-4 Technical Report". Marzo de 2023. <https://arxiv.org/abs/2303.08774>

OpenAI. "Introducing ChatGPT". Blog de OpenAI, 30 de noviembre de 2022. Accedido el 18 de mayo de 2024. <https://www.openai.com/blog/chatgpt>

OpenAI. "Learning to Reason with LLMs". Publicado el 12 de septiembre de 2024.
<https://openai.com/index/learning-to-reason-with-langs/>

OpenAI. OpenAI o1 System Card. 5 de diciembre de 2024. <https://cdn.openai.com/o1-system-card-20241205.pdf>

Organización Internacional de Normalización. "Normas". Accedido el 26 de junio de 2024.
<https://www.iso.org/standards.html>

Ostermann, Adrian, y Niklas Jooß. "Interoperability: definition, evaluation and application". FfE Munich, 16 de noviembre de 2022.

<https://www.ffe.de/en/veroeffentlichungen/Interoperability-definition-evaluation-and-application>

Oubělický, Richard. "Europe and AI: Causes and Implications of Europe Losing Ground in the Race for AI (Part I)". Security Outlines, 13 de marzo de 2024. <https://securityoutlines.cz/europe-and-ai-causes-and-implications-of-europe-losing-ground-in-the-race-for-ai-part-i/>

Oubělický, Richard. "Europe and AI: Causes and Implications of Europe Losing Ground in the Race for AI": 35-50.

Petkova, Bilyana. "Privacy as Europe's First Amendment". European Law Journal 25, no. 2 (2019): 140-154. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333937

Pichai, Sundar, y Demis Hassabis. "Our next-generation model: Gemini 1.5". The Keyword, Google, 15 de febrero de 2024. <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note>

Poder Judicial de Costa Rica. "Agenda de Corte Plena – Lunes 6 de noviembre de 2023". 6 de noviembre de 2023. Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/agenda-de-corte-plena-lunes-6-de-noviembre-de-2023?catid=8&Itemid=409#:~:text=para%20la%20promoci%C3%B3n%20responsables%20de,919>

Poder Judicial de Costa Rica. "Analizan los retos y oportunidades post pandemia de las Tics en la Administración de Justicia". Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/analizan-los-retos-y-oportunidades-post-pandemia-de-las-tics-en-la-administracion-de-justicia?catid=8&Itemid=409#:~:text=%E2%80%9CSe%20destaca%20que%20la%20intel>

[igencia, Uni%C3%B3n%20Europea%E2%80%9D%20detall%C3%B3n%20Bujosa%20Vadell](#)

Poder Judicial de Costa Rica. "Justicia Abierta". Recuperado de: <https://servicios.poder-judicial.go.cr/index.php/funcionamiento-y-los-programas-pj/41-justicia-abierta>

Poder Judicial de Costa Rica. "Juzgados Especializados de Cobro de San José se preparan para trabajar con Inteligencia Artificial". 7 de junio de 2023. Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/juzgados-especializados-de-cobro-de-san-jose-se-preparan-para-trabajar-con-inteligencia-artificial?catid=8&Itemid=409#:~:text=Juzgados%20Especializados%20de%20Cobro%20de,a%20ponerse%20en%20marcha>

Poder Judicial de Costa Rica. "Novedosa herramienta de Inteligencia Artificial se aplica en mejora de la protección de datos". Presentado el 20 de marzo de 2024. Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/novedosa-herramienta-de-inteligencia-artificial-se-aplica-en-mejora-de-la-proteccion-de-datos?catid=8&Itemid=409>

Poder Judicial de Costa Rica. "Planificación estratégica dirigida a la modernización del Poder Judicial". Recuperado de: <https://pj.poder-judicial.go.cr/index.php/component/content/article/planificacion-estrategica-dirigida-a-la-modernizacion-del-poder-judicial?catid=8&Itemid=409#:~:text=Con%20el%20apoyo%20de%20la,informaci%C3%B3n%20en%20el%20quehacer%20institucional>

Poder Judicial de Costa Rica. "Poder Judicial implementa inteligencia artificial para disminuir circulante en materia cobratoria". Recuperado de: <https://pj.poder-judicial.go.cr>

judicial.go.cr/index.php/component/content/article/poder-judicial-implementa-inteligencia-artificial-para-disminuir-circulante-en-materia-cobratoria?catid=8&Itemid=409 Predictice. "Accédez à toute l'information juridique". Predictice. Consultado el 22 de mayo de 2024. <https://www.predictice.com>

Radford, Alec, Karthik Narasimhan, Tim Salimans e Ilya Sutskever. "Improving Language Understanding by Generative Pre-Training". OpenAI, 2018. <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>

Rafailov, Rafael, Chris Schutte, Michael Noukhovitch, Scott Lundberg, Jesse Mu, y Jacob Steinhardt. "Direct Preference Optimization: Your Language Model is Secretly a Reward Model". arXiv, 13 de diciembre de 2023. Stanford University. <http://arxiv.org/abs/2305.18290>.

Rawls, John. La justicia como equidad: Una reformulación. Editado por Erin Kelly. Barcelona: Paidós, 2001. Recuperado de: <https://www.terrileyasociados.com.ar/post/john-rawls-la-justicia-como-equidad-una-reformulacion-a-cargo-de-erin-kelly-paidos.pdf>

Re, Richard M., y Alicia Solow-Niederman. "Developing Artificially Intelligent Justice". Stanford Technology Law Review 22 (2019): 242-289. https://law.stanford.edu/wp-content/uploads/2019/08/Re-Solow-Niederman_20190808.pdf

Reinsch, William Alan, Matthew Schleich, y Thibault Denamiel. "Insight into the U.S. Semiconductor Export Controls Update". Center for Strategic and International Studies (CSIS), 20 de octubre de 2023. <https://www.csis.org/analysis/insight-us-semiconductor-export-controls-update>

Reiter, R. "A Logic for Default Reasoning". Artificial Intelligence 13, no. 1-2 (1980): 81-132. <https://www.sciencedirect.com/science/article/pii/0004370280900144>

Remus, Dana, and Frank Levy. "Can Robots Be Lawyers? Computers, Lawyers, and the Practice of Law." Georgetown Journal of Legal Ethics 30, no. 3 (2017): 501–58.
<https://dx.doi.org/10.2139/ssrn.2701092>

Roa Avella, Marcela del Pilar, y Jesús Eduardo Sanabria-Moyano. "Uso del algoritmo COMPAS en el proceso penal y los riesgos a los derechos humanos". Artículo producto del proyecto INV DER 3159 "Inteligencia Artificial: retos y riesgos de los Derechos Humanos en el Sistema Penal", financiado por la Vicerrectoría de Investigaciones de la Universidad Militar Nueva Granada, convocatoria Proyectos de Investigación Científica vigencia 2020.
<https://www.google.co.cr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiVvNXh3qKGAxUFVTABHUU3A1UQFnoECBQQAQ&url=https%3A%2F%2Fdialnet.unirioja.es%2Fdescarga%2Farticulo%2F8438795.pdf&usg=AOvVaw3xLxCMaUz1OcM7j ug0Zmab&opi=89978449>

Rodríguez Solís, Marisel. "OIJ apuesta por la inteligencia artificial". Poder Judicial, 23 de enero de 2023. Recuperado de: <https://pjenlinea3.poder-judicial.go.cr/biblioteca/uploads/Archivos/Articulo/OIJ%20apuesta%20por%20la%20inteligencia%20artificial.pdf>

Rosenblatt, Frank. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". Psychological Review 65, no. 6 (1958): 386-408. 1959-09865-001.pdf (apa.org).

Rumelhart, David E., Geoffrey E. Hinton, y Ronald J . Williams. "Learning representations by back-propagating errors". Nature 323, no. 6088 (1986): 533-536.
<https://www.nature.com/articles/323533a0>

Russell, Stuart J., y Peter Norvig. Artificial Intelligence: A Modern Approach. 3^a ed. Upper Saddle River, NJ: Pearson Education, Inc., 2010.
https://people.engr.tamu.edu/guni/csce421/files/AI_Russell_Norvig.pdf

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Resolución n.º 3454–2012 del 9 de marzo de 2012.

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Resolución n.º 6240–1993 del 26 de noviembre de 1993.

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Resolución n.º 12046–2012 del 1 de agosto de 2012.

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Resolución n.º 136–2003 del 15 de enero de 2003.

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Resolución n.º 598–1990 del 30 de mayo de 1990.

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Resolución n.º 10734–2004 del 29 de septiembre de 2004.

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Resolución n.º 14519–2005 del 21 de octubre de 2005.

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Resolución N° 16787-2018, Expediente 18-009250-0007-CO del 5 de octubre de 2018.

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Voto No. 8390-97 del 9 de diciembre de 1997.

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Voto N.º 4849-2009 del 20 de marzo de 2009.

Sala Constitucional de la Corte Suprema de Justicia de Costa Rica. Voto N.º 5790-99 del 11 de agosto de 1999.

Samuel, Arthur L. "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development 3, no. 3 (1959): 210-229. <https://doi.org/10.1147/rd.33.0210>.

Schaeffer, Rylan, Brando Miranda, y Sanmi Koyejo. "Are Emergent Abilities of Large Language Models a Mirage?". Computer Science, Stanford University, mayo de 2023. arXiv, 2304.15004.pdf (arxiv.org).

Schauer, Frederick. "The Role of the Text: Easy Cases". En Methods of Constitutional Interpretation.

http://fs2.american.edu/dfagel/www/Class%20Readings/Schauer/Schauer%20Easy%20Cases%20Only_CleanedUp.pdf

Schwab, Klaus. La cuarta revolución industrial. Ginebra: Foro Económico Mundial, 2016. Edición en español, México: Penguin Random House Grupo Editorial, S. A. de C. V., 2017. Recuperado de: <https://economiapoliticaenam.wordpress.com/wp-content/uploads/2020/05/klaus-schwab.la-4c2b0-rev.-industrial-2.pdf>

Shannon, Claude E. "Programming a Computer for Playing Chess". The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 41, no. 314 (1950): 256-75. <https://doi.org/10.1080/14786445008521796>

Shi, Changqing, Tania Sourdin, y Bin Li. "The Smart Court – A New Pathway to Justice in China?". International Journal for Court Administration 12, no. 1 (2021): 4-19. <https://storage.googleapis.com/jnl-up-j-ijca-files/journals/1/articles/367/submission/proof/367-1-1754-2-10-20210311.pdf>

Shortliffe, Edward Hance. Computer-Based Medical Consultations: MYCIN. New York: American Elsevier Publishing Company, Inc., 1976.

<https://momot.rs/d3/y/1710673429/100/u/annas-archive-ia-2023-06-lcpdf/c/computerbasedmed0000shor.pdf~/HSU8XEUovUQ2oN-z1qecHQ/Computer-based%20medical%20consultations%2C%20MYCIN%20--%20Shortliffe%2C%20Edward%20Hance%20--%201976%20--%20New%20York%3A%20Elsevier%20--%209780444001795%20--%20e266cab310354dd26f341b4e1713fed4%20--%20Anna%20%20%20Archive.pdf>

Simon, Herbert A. The Shape of Automation for Men and Management. Nueva York: Harper & Row, Publishers, 1965. <https://ebin.pub/download/the-shape-of-management-for-men-and-management.html>

Simoncini, Andrea. "L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà". BioLaw Journal – Rivista di BioDiritto 1, no. 1 (2018): 81. Recuperado de: https://www.academia.edu/94336762/L_algoritmo_incostituzionale_intelligenza_artificiale_e_il_futuro_delle_libert%C3%A0

"State v. Loomis. La Corte Suprema de Wisconsin exige advertencia previa al uso de evaluaciones algorítmicas de riesgo en la determinación de sentencias (comentario sobre 881 N.W.2d 749 [Wis. 2016])". Harvard Law Review 130, no. 5 (marzo de 2017). Disponible en: <https://harvardlawreview.org/print/vol-130/state-v-loomis/>

Sutskever, Ilya. "Open-Source vs. Closed-Source AI". En Inside OpenAI [Entire Talk], entrevistado por Ravi Belani. Stanford eCorner, 26 de abril de 2023. <https://ecorner.stanford.edu/clips/open-source-vs-closed-source-ai/>

Susskind, Richard. *Tomorrow's Lawyers: An Introduction to Your Future*. 2nd ed. Oxford: Oxford University Press, 2017. <https://pdfroom.com/books/tomorrows-lawyers-an-introduction-to-your-future/NpgpZJQe5jr/download>

Tamkin, Alex, Miles Brundage, Jack Clark, y Deep Ganguli. "Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models". arXiv:2102.02503v1 [cs.CL], 4 de febrero de 2021. 2102.02503.pdf (arxiv.org).

Tegmark, Max. Life 3.0: Being Human in the Age of Artificial Intelligence. New York: Alfred A. Knopf, 2017. Life 3.0: Being Human in the Age of Artificial Intelligence (readyforai.com).

Tegmark, Max. Vida 3.0: Ser humano en la era de la Inteligencia Artificial. España: Editorial Taurus, 2018. (PDF) Vida 3. | Doroteo Arango - Academia.edu.

Templeton, Adly, Geoffrey Irving, Ethan Pérez, Gregory Kobren, Tao Lin, y Owain Evans. "Scaling Monosemantics: Extracting Interpretable Features from Claude 3 Sonnet". Anthropic, 21 de mayo de 2024. <https://transformer-circuits.pub/2024/scaling-monosemantics/index.html>

Templeton, Adly, Jamie Clapman, Reuben Aronson, Ronan Toal, y Danielle Grinspan. "Scaling Monosemantics: Extracting Interpretable Features from Claude 3 Sonnet". Anthropic, 21 de mayo de 2024. <https://transformer-circuits.pub/2024/scaling-monosemantics/index.html>

The Law Society Commission on the Use of Algorithms in the Justice System. Algorithms in the Criminal Justice System. Londres: The Law Society of England and Wales, junio de 2019. <https://www.lawsociety.org.uk/topics/research/algorithm-use-in-the-criminal-justice-system-report>

Thomson Reuters. "Comply or Explain Approach". Practical Law Glossary. Resource ID 8-107-5967, 2024. [https://uk.practicallaw.thomsonreuters.com/8-107-5967?transitionType=Default&contextData=\(sc.Default\)&firstPage=true](https://uk.practicallaw.thomsonreuters.com/8-107-5967?transitionType=Default&contextData=(sc.Default)&firstPage=true)

Torrealba Navas, Federico. Principios del Derecho Privado. San José, Costa Rica: IJ Editores, Librería y Editorial Juricentro, abril 2021.

Tribunal Contencioso Administrativo y Civil de Hacienda. Voto No. 2024-4230 del 1 de julio de 2024.

Tribunal de Justicia de la Unión Europea. Sentencia de 13 de diciembre de 1989, Salvatore Grimaldi contra Fonds des maladies professionnelles. C-322/88, ECLI:EU:C:1989:646. EUR-Lex. <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A61988CJ0322>

Tribunal de Justicia de la Unión Europea. Sentencia de 21 de septiembre de 1983, Deutsche Milchkontor GmbH y otros contra República Federal de Alemania. Asuntos acumulados 205 a 215/82, ECLI:EU:C:1983:233.

Tribunal de Justicia de la Unión Europea. Sentencia del Tribunal General (Sala Séptima ampliada) de 12 de junio de 2014, Intel Corp. contra Comisión Europea. Asunto T-286/09, ECLI:EU:T:2014:547. <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62009TJ0286>

Tribunal Europeo de Derechos Humanos. Caso Baka contra Hungría, sentencia del 23 de junio de 2016. Recuperado de: [https://hudoc.echr.coe.int/eng#%22itemid%22:\[%22001-144139%22\]}](https://hudoc.echr.coe.int/eng#%22itemid%22:[%22001-144139%22]})

Tribunal Europeo de Derechos Humanos. Caso Maktouf y Damjanović contra Bosnia y Herzegovina, sentencia del 18 de julio de 2013. Recuperado de: [https://hudoc.echr.coe.int/fre#%22itemid%22:\[%22002-4870%22\]}](https://hudoc.echr.coe.int/fre#%22itemid%22:[%22002-4870%22]})

Turmo Borras, Jordi. "Herramientas de extracción de información: redes neuronales". En Extracción de información en textos escritos en español. Tesis doctoral, Universidad de Barcelona, 2000. http://deposit.ub.edu/dspace/bitstream/2445/35334/5/3.CAPITULO_2.pdf

Turing, Alan. "Maquinaria computacional e inteligencia". Traducido por Cristóbal Fuentes Barassi. *Philosophy* 36, no. 136 (1950): 433-460.
<https://doi.org/10.1017/S0031819100060491>

UNESCO. "Herramientas para un uso ético: Jueces de América Latina y el Caribe se capacitan en Inteligencia Artificial y Estado de Derecho". Noticia, 15 de noviembre de 2023. Recuperado de: <https://es.unesco.org/news/herramientas-para-un-uso-etico-jueces-de-america-latina-y-caribe-capacitan-en-inteligencia-artificial>

Unión Europea. "Carta de los Derechos Fundamentales de la Unión Europea". Diario Oficial de la Unión Europea, C 326, 26 de octubre de 2012. <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:12012P/TXT&from=EN>

Unión Europea. Versión Consolidada del Tratado de Funcionamiento de la Unión Europea, art. 288, 2012 O.J. C 326/47. <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A12012E%2FTXT>

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gómez, Lukasz Kaiser, y Illia Polosukhin. "Attention Is All You Need". Advances in Neural Information Processing Systems, 2017. [1706.03762] Attention Is All You Need (arxiv.org).

Velázquez, Hugo José Francisco. "Esclareciendo el concepto de lógica deontológica". Revista Andamios 18, no. 45 (enero-abril 2021): 459-482.
https://uacm.edu.mx/portals/5/num45/19_A_Esclareciendo.pdf

Vollmer, Andrew N., y John Byron Sandage. "The Wood Pulp Case". International Law 23, no. 3 (1989): 721-732. <https://scholar.smu.edu/tl/vol23/iss3/9>

Walker, Stephen M., II. "What is supervised fine-tuning?". Klu.AI blog. <https://www.klu.ai/what-is-supervised-fine-tuning>

Winograd, Terry. "Understanding Natural Language". *Cognitive Psychology* 3, no. 1 (1972): 1-191. Massachusetts Institute of Technology, Cambridge, MA.
<https://www.sciencedirect.com/science/article/pii/0010028572900023?via%3Dihub>

WIPO. WIPO Technology Trends 2019: Artificial Intelligence. Ginebra: Organización Mundial de la Propiedad Intelectual, 2019.
https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf

Zagrebelsky, Gustavo. El derecho dúctil: Ley, derechos, justicia. Traducido por Marina Gascón. Madrid: Editorial Trotta, 2011. Recuperado de:
https://www.academia.edu/4980303/155026921_El_Derecho_Ductil_Gustavo_Zagrebelsky_pdf

Zhabina, Alena. "Cómo la IA de China está automatizando el sistema legal". DW, 20 de enero de 2023. <https://p.dw.com/p/4MUY0>.