Inteligencia artificial, opacidad algorítmica y decisiones judiciales: necesidad de establecer una moratoria

Artificial Intelligence, Algorithmic Opacity, and Judicial Decision-Making: The Need to Set a Moratorium

Inteligência Artificial, Opacidade Algorítmica E Decisões Judiciais: Necessidade De Estabelecer Uma Moratória

Gabriel Ravelo Franco[1]
Juan Pablo Sarmiento Erazo[2]
Jesús E. Sanabria Moyano [3]

Recibido: 16 de diciembre de 2025
Aprobado: 2 de febrero de 2026
Publicado: 27 de abril de 2026

Cómo citar este artículo:
Ravelo Franco, G., Sarmiento Erazo, J.P. y Sanabria Moyano, J.E. (2026) Inteligencia artificial, opacidad algorítmica y decisiones judiciales: necesidad de establecer una moratoria. Especial DIXI -RI/INS 2026 | Interdisciplinariedad + Innovación en las ciencias jurídicas, 1-22.
DOI: https://doi.org/10.16925/2382-4220.2026.03.02

Artículo de investigación. https://doi.org/10.16925/2382-4220.2026.03.02

1 Maestro en Derecho Penal. Profesor y doctorando en Derecho de la Escuela de Posgrado, Universidad Continental, Lima, Perú.
Correo electrónico: gravelo@continental.edu.pe

2 Maestro y Doctor en Derecho. Profesor de la Universidad de la Sabana, Colombia.
Correo electrónico: juan.sarmiento3@unisabana.edu.co

3 Magíster en Derecho Público. Profesor de la Universidad Militar de Nueva Granada, Colombia.
Correo electrónico: jesus.sanabria@unimilitar.edu.co

Resumen

El estudio examina la viabilidad del uso de sistemas de inteligencia artificial (IA) en decisiones jurisdiccionales a partir de la problemática estructural de la opacidad algorítmica (OA). El trabajo contrasta las posturas que admiten niveles de opacidad funcionalmente tolerables con aquellas que sostienen que toda ininteligibilidad afecta la legitimidad democrática, la justicia epistémica y los derechos fundamentales. Asimismo, integra aportes de la XAI y del análisis probatorio para mostrar que las explicaciones disponibles no permiten reconstruir razonamientos, hechos ni vínculos causales indispensables en el proceso judicial. A través de una revisión exhaustiva de literatura reciente y del análisis de estándares normativos, filosóficos y procesales, se demuestra que las limitaciones técnicas, epistémicas, sociotécnicas e institucionales de los modelos actuales impiden garantizar condiciones mínimas de explicabilidad, auditabilidad y control racional. A la luz de estos hallazgos y del estándar interamericano sobre el derecho a comprender las razones de una decisión, se propone una moratoria precautoria que suspenda la incorporación de sistemas de IA en el ámbito jurisdiccional hasta que se validen estándares robustos de explicabilidad mediante procesos deliberativos interinstitucionales.

Palabras clave: explicabilidad; debido proceso; control jurisdiccional; riesgo epistémico; motivación judicial.¡

Abstract

This study examines the feasibility of employing artificial intelligence systems in judicial decision-making in light of the structural challenges posed by algorithmic opacity. It contrasts approaches that regard certain levels of functional opacity as tolerable with those that contend that any form of inintelligibility undermines democratic legitimacy, epistemic justice, and fundamental rights. The analysis also incorporates insights from XAI and evidentiary theory to show that existing explanation methods do not enable the reconstruction of the reasoning, facts, or causal links essential for judicial processes. Through an exhaustive review of recent scholarship and an assessment of normative, philosophical, and procedural standards, the study demonstrates that the technical, epistemic, sociotechnical, and institutional limitations of current models preclude meeting the minimum requirements of explainability, auditability, and rational control. In view of these findings and the Inter-American standard affirming the right to understand the reasons underlying a judicial decision, the article proposes a precautionary moratorium on the use of AI in the judicial domain until robust explainability standards are validated through inter-institutional deliberative processes.

KeywordsExplainability; due process; judicial scrutiny; epistemic risk; judicial reasoning

Resumo

O estudo examina a viabilidade do uso de sistemas de inteligência artificial (IA) em decisões jurisdicionais a partir da problemática estrutural da opacidade algorítmica (OA). O trabalho contrapõe as posições que admitem níveis de opacidade funcionalmente toleráveis àquelas que sustentam que toda ininteligibilidade compromete a legitimidade democrática, a justiça epistêmica e os direitos fundamentais. Ademais, integra aportes da XAI e da análise probatória para demonstrar que as explicações atualmente disponíveis não permitem reconstruir os raciocínios, os fatos nem os vínculos causais indispensáveis ao processo judicial. Por meio de uma revisão abrangente da literatura recente e da análise de padrões normativos, filosóficos e processuais, demonstra-se que as limitações técnicas, epistêmicas, sociotécnicas e institucionais dos modelos atuais impedem a garantia de condições mínimas de explicabilidade, auditabilidade e controle racional. À luz desses achados e do padrão interamericano relativo ao direito de compreender as razões de uma decisão, propõe-se uma moratória precautória que suspenda a incorporação de sistemas de IA no âmbito jurisdicional até que sejam validados padrões robustos de explicabilidade por meio de processos deliberativos interinstitucionais.

Palavras-Chave Explicabilidade; devido processo; controle jurisdicional; risco epistêmico; fundamentação judicial.

Introducción

Una de las principales preocupaciones actuales relativas a la Inteligencia Artificial (IA) es la OA, es decir, la dificultad para conocer, comprender o auditar el funcionamiento interno de los algoritmos que orientan o determinan resoluciones judiciales. Esta opacidad puede adoptar diversas formas —desde la complejidad técnica hasta restricciones legales o comerciales— y plantea un desafío a los principios de transparencia, rendición de cuentas y control jurisdiccional.

Diversos estudios han abordado la noción de OA. Facchini y Termine (2022) exploran diferentes taxonomías propuestas, como la de Burrell, quien distingue tres manifestaciones: la opacidad derivada del secreto corporativo o estatal, la asociada a la falta de alfabetización técnica, y aquella que resulta del funcionamiento mismo de los algoritmos en contextos reales de aplicación; también abordan la de Creel, quien propone una tipología basada en niveles de abstracción: opacidad de ejecución (run opacity), opacidad estructural y OA, las cuales aluden respectivamente a la comprensión del proceso físico, del código y de la lógica general del sistema.

Dicho esto, las arquitecturas de aprendizaje profundo, en particular las redes neuronales artificiales (ANNs), ilustran con claridad este problema. A pesar de ser eficaces para tareas predictivas, su funcionamiento interno resulta, en muchos casos, ininteligible incluso para sus propios programadores. Como ha advertido Carabantes (2020), estas estructuras replican la funcionalidad del neocórtex y predicen comportamientos humanos a partir de fragmentos de información, pero lo hacen mediante operaciones que no se articulan en lógica comprensible: sin palabras, sin argumentos, sin razonamientos. Su capacidad de predicción constituye una forma de poder técnico que compromete la posibilidad de control y de rendición de cuentas. Con ello, la sociedad contemporánea incurre en una forma de racionalidad instrumental que privilegia la eficacia sobre la inteligibilidad, y se aproxima, como diría Horkheimer, a una forma moderna de mito disfrazado de ciencia.

En esta lógica, Cantarini (2024) cuestiona si puede hablarse de una legitimación social de los sistemas algorítmicos cuando las personas no son conscientes de estar sometidas a ellos, y cuando además carecen de mecanismos de control o rendición de cuentas. La autora sugiere que no estamos ante instituciones consolidadas, sino más bien frente a un proceso de institucionalización de los algoritmos. Y si lo que deseamos es contar con estructuras sociales que garanticen la convivencia social, cabe dudar que estos sistemas —que plantean amenazas documentadas a los derechos fundamentales— puedan recibir tal calificación.

En esta misma línea, Grant et al. (2025) argumentan que existen deberes morales hacia los sujetos de las decisiones que justifican, en muchos casos, rechazar el uso

de sistemas basados en IA opaca. Estos deberes incluyen la obligación de emplear reglas de inferencia y decisión permisibles, así como el deber de brindar consideración agencial a las personas afectadas. La imposibilidad de verificar si los sistemas cumplen con estos estándares, o incluso la certeza de que no pueden cumplirlos, socava la legitimidad del proceso. Aun cuando los niveles de transparencia exigibles varían según el contexto, esta perspectiva refuerza la necesidad de evaluar la delegación de decisiones a sistemas que no están sujetos a las mismas restricciones éticas que los operadores humanos.

Este artículo aborda una posible intersección entre la OA y el principio precautorio. Si bien el principio de precaución ha ganado mayor peso en conflictos ambientales, la experiencia comparada la ha extendido a otros escenarios de incertidumbre. En este sentido, podemos constatar los aportes de Ruiz Jarabo (2005) que considera que en su esencia el principio de precaución es un principio que otorga a las diferentes autoridades y poderes públicos la facultad de adoptar las medidas preventivas cuando exista incertidumbre científica (Ruiz, 2005). La estricta aplicación del principio de precaución implica que "si existen indicios de daño (en contraste con la certeza), deberá presumirse que la actividad es dañina, hasta que de manera concluyente se pruebe lo contrario" (Riechmann, & Tickner, 2002). Luego, como se procederá a demostrar en este artículo, el riesgo y compromiso a garantías fundamentales, así como la opacidad de los algoritmos en decisiones judiciales sobre derechos de las personas, se encuentran en un marco de incertidumbre que permitiría activar el mentado principio.

El presente artículo parte de la siguiente interrogante: ¿Debe establecerse una moratoria en el uso de sistemas automatizados para la toma de decisiones judiciales, con fundamento en el principio precautorio, ante los riesgos que plantea la OA para la garantía de los derechos fundamentales y la tutela jurisdiccional efectiva? Así, el objetivo general consiste en demostrar si resulta necesario imponer una moratoria en el uso de decisiones judiciales automatizadas, con base en el principio precautorio, como medida preventiva frente a los riesgos que la OA representa para los derechos fundamentales y la tutela jurisdiccional efectiva.

Materiales y métodos

El trabajo consiste en una propuesta de moratoria que toma como insumos una revisión de literatura ejecutada en Scopus el 31 de julio de 2025 bajo la siguiente ecuación: (TITLE-ABS-KEY (opacity)) AND (burrell) AND (ai). Dicha búsqueda está centrada en la opacidad de la IA en función de uno de los referentes en la materia: Burrell. Los resultados iniciales de la búsqueda fueron 154 resultados, de

e-ISSN 2382-4220 / Vol.28, n°. 3 / Especial DIXI -RI/INS 2026
Universidad Cooperativa de Colombia

los cuales 10 fueron excluidos a partir de la revisión de títulos y resúmenes. Así, se obtuvo una base preliminar de 144 artículos. Luego, de la revisión de contenido, se identificaron los trabajos más relevantes en función del propósito de este estudio, tomando como punto de partida la reiteración a las expresiones opacity y XAI, los cuales se presentan en la sección de resultados.

Definición y taxonomía de la OA

A continuación, se presenta la síntesis derivada de una revisión de literatura conforme a lo descrito en la sección de Materiales y Métodos.

Tabla 1

Definición y taxonomía de la opacidad algorítmica

Autor (año)	Aporte del autor
Mann et al. (2023)	Identifican ocho fuentes de opacidad, organizadas en tres categorías: arquitectónica (complejidad y representaciones extranjeras), analítica (habilidades ausentes, herramientas inexistentes y recursos insuficientes) y sociotécnica (dependencia epistémica, conocimiento perdido y ocultamiento intencional). Esta clasificación proporciona una guía estructurada para seleccionar estrategias contextuales de intervención.
Facchini & Termine (2022)	Proponen una taxonomía compuesta por tres dimensiones: opacidad de acceso, de vínculo y semántica. Cada una alude a las barreras para comprender el funcionamiento interno, establecer correspondencias con fenómenos del mundo y atribuir significado al contenido y razonamiento del modelo. Este enfoque contextualiza la opacidad según el usuario y el propósito.
Freiman et al. (2025)	Introducen una distinción entre opacidad epistémica y no epistémica, señalando que la primera afecta la posibilidad de conocer cómo y por qué se produce una salida, mientras que la segunda responde a condiciones como el uso de proxies, normas o datos deficientes. Esta diferenciación permite esclarecer la fuente de l problema y proponer soluciones más ajustadas.
Petrolo et al. (2023)	Desarrollan el modelo IPO, definiendo la opacidad epistémica de algoritmos según la ausencia de justificación epistémica sobre sus entradas, procedimientos o salidas. Su enfoque combina una base epistemológica con lógica modal para analizar estas condiciones, proporcionando un marco formal para evaluar la transparencia en sistemas de aprendizaje automático.

Autor (año)	Aporte del autor
Gerdes (2021)	Explora la OA desde una perspectiva ética, distinguiendo entre opacidad técnica, deliberada y emergente. Argumenta que cada tipo plantea distintos riesgos para la autonomía y el juicio moral de los usuarios, e insiste en la necesidad de considerar la agencia humana al diseñar mecanismos de transparencia en entornos automatizados.
Kroll (2018)	Examina la OA como problema de gobernanza. Vincula la posibilidad de auditoría y control legal con el diseño técnico de los sistemas y propone un enfoque que combina mecanismos de rendición de cuentas ex ante y ex post, articulando soluciones desde el derecho administrativo y la informática para reducir la opacidad en entornos automatizados.
Kaas (2024)	Distingue entre opacidad epistémica, metodológica y política. Propone abordajes diferenciados para cada tipo: técnicas explicativas y visuales para la primera, análisis de diseño y experimentación para la segunda, y estudios críticos sobre gobernanza algorítmica para la tercera. Esta clasificación promueve una comprensión plural de la opacidad en contextos sociotécnicos.
Gorwa et al. (2020)	Articulan una taxonomía centrada en la gobernanza algorítmica, que distingue entre opacidad técnica, institucional y normativa. Esta diferenciación permite analizar no solo la comprensibilidad técnica del algoritmo, sino también los marcos organizacionales y legales que inciden en su transparencia y control.
Lee (2025)	Propone un enfoque pragmático de la OA centrado en la experiencia del usuario. Clasifica la opacidad en funcional, relacional y estructural, destacando que cada tipo responde a obstáculos específicos en la interacción humana con sistemas automatizados, lo cual permite identificar necesidades diferenciadas de intervención.
Alvarado (2023)	Examina la OA desde la justicia epistémica, señalando cómo ciertas prácticas de diseño y uso de sistemas automatizados pueden excluir saberes situados. Propone una aproximación crítica que visibilice asimetrías cognitivas, especialmente en contextos institucionales, promoviendo una comprensión de la opacidad como fenómeno distribuido y relacional.
Carabantes (2020)	Resalta la utilidad de la propuesta de Burrell sobre las tres opacidades —intencional, técnica y interpretativa—, para analizar la falta de transparencia como fenómeno estructurado. Destaca que estas formas permiten distinguir entre límites cognitivos, diseño deliberado y dificultades de interpretación, lo que contribuye a una comprensión más matizada de los obstáculos en la inteligibilidad de los sistemas algorítmicos.
Lo (2024)	Desde una perspectiva filosófica eetoma las tres formas propuesta Burrell (intencional, técnica e interpretativa). Sostiene que cada una refleja una dimensión distinta de la distancia cognitiva entre usuarios y sistemas, y propone abordajes específicos para reducir dicha distancia mediante estrategias de diseño epistémico y reconstrucción conceptual de los procesos automatizados.

e-ISSN 2382-4220 / Vol.28, n°. 3 / Especial DIXI -RI/INS 2026
Universidad Cooperativa de Colombia

Autor (año)	Aporte del autor
Knoks & Raleigh (2022)	Abordan la opacidad de redes neuronales profundas retomando la noción de Burrell sobre cómo operan los algoritmos a gran escala. Consideran que la explicación en XAI debe evaluarse según estándares filosóficos de comprensión y modelos científicos, y advierten que los métodos actuales pueden generar una falsa apariencia de inteligibilidad sin una conexión formal verificable con el sistema original.
Langer & König (2023)	Introducen un enfoque multiactor que distingue tres fuentes de OA: técnica (relacionada con la complejidad del sistema), epistémica (derivada de la falta de alfabetización técnica) e intencional (vinculada a decisiones deliberadas de mantener opaco el sistema). Analizan sus efectos diferenciados sobre usuarios, afectados, desarrolladores, implementadores y reguladores.
Bjerring et al. (2025)	Proponen un modelo deliberativo de legitimidad algorítmica que exige condiciones de inteligibilidad pública. Argumentan que la opacidad impide el escrutinio racional necesario para que los sistemas algorítmos sean aceptables en contextos democráticos. Defienden que solo los sistemas explicables pueden cumplir los estándares de justificación requeridos por el ideal deliberativo.
Lu (2022)	Analiza la OA como barrera para la explicabilidad desde una perspectiva de derechos. Plantea que la falta de comprensión del funcionamiento algorítmico compromete el derecho a la información y la autodeterminación informativa, especialmente en decisiones automatizadas que afectan a individuos sin posibilidad de apelación efectiva.
Kumar et al. (2024)	Proponen una evaluación funcional de la opacidad basada en contextos de uso. Argumentan que no toda opacidad es problemática si los resultados son fiables y auditables por terceros. Introducen el concepto de “opacidad aceptable” como equilibrio entre inteligibilidad técnica, eficiencia operativa y control institucional.
Hatherley (2025)	Sostiene que los sistemas opacos, al ser inaccesibles para el escrutinio público, consolidan una forma de autoridad no verificable que afecta la distribución del conocimiento en sociedades democráticas. Propone una teoría de la explicabilidad como condición de justicia epistémica.
Boge (2022)	Distingue *entre entender un modelo* y *entender con un modelo*, y sostiene que muchos sistemas de aprendizaje automático no permiten lo segundo. La opacidad no radica solo en la complejidad interna, sino en la incapacidad de usar el modelo como herramienta cognitiva para comprender el mundo, lo que limita su valor explicativo más allá de la predicción.
Lasbleiz & Milkes (2021)	Opacidad intencional: Oculta el funcionamiento del algoritmo para proteger la información Opacidad analfabeta: El sistema sólo puede ser comprendido por quienes prueba utilizar y entender el código Opacidad intrínseca: un proceso tan complejo que no es fácil de comprender.

Nota general. Elaboración propia.

Un primer consenso identificado es el relativo a la OA, en el sentido de que no se limita a un problema técnico, sino que involucra dimensiones epistémicas y sociotécnicas. Mann et al. (2023), al clasificar las fuentes de opacidad en arquitectónica, analítica y sociotécnica, anticipan una comprensión integral que luego será retomada por Kaas (2024) al distinguir entre opacidad epistémica, metodológica y política. Ambos enfoques subrayan que los obstáculos cognitivos, los límites del diseño y los contextos institucionales interactúan en la producción de ininteligibilidad.

Asimismo, existe consenso relativo a que la explicabilidad no equivale a transparencia plena. Freiman et al. (2025) y Petrolo et al. (2023) destacan que la comprensión de un modelo no garantiza su justificabilidad, y que la ausencia de razones verificables sobre las relaciones entre entradas, procesos y salidas constituye el núcleo de la opacidad epistémica. En la misma línea, Gerdes (2021) y Kroll (2018) defienden la necesidad de mecanismos de rendición de cuentas —éticos o jurídicos— que complementen las soluciones puramente técnicas.

Otro punto de convergencia reside en la preocupación por la agencia humana y el control democrático. Alvarado (2023) plantea que la opacidad perpetúa asimetrías cognitivas y exclusiones epistémicas, mientras que Gerdes (2021) advierte que la opacidad deliberada o emergente debilita la responsabilidad moral. Ambos coinciden con Kroll (2018), quien sostiene que la gobernanza algorítmica debe articular controles ex ante y ex post para restablecer el equilibrio entre eficacia técnica y legitimidad pública.

Sin mebargo, los enfoques también presentan divergencias en la manera de conceptualizar el núcleo de la opacidad. Para Mann et al. (2023), se trata de un fenómeno estructural que puede ubicarse en diferentes niveles de un sistema, mientras que para Facchini y Termine (2022) la clave reside en las barreras comunicativas: la dificultad de acceso, vínculo o interpretación semántica entre el modelo y el usuario. Freiman et al. (2025) proponen un criterio normativo —epistémico versus no epistémico— que no es estrictamente técnico ni fenomenológico, sino evaluativo. Petrolo et al. (2023) añaden una formalización lógica (modelo IPO) que, si bien sistematiza la relación entre entradas y salidas, deja sin resolver los aspectos políticos e institucionales.

También existen divergencias respecto a la finalidad de la clasificación. Mann et al. (2023) y Kaas (2024) buscan ofrecer tipologías analíticas para ubicar los focos de ininteligibilidad; Gerdes y Alvarado, en cambio, abordan la opacidad como un problema ético y de justicia cognitiva; y Kroll (2018) la concibe desde la teoría de la gobernanza y la rendición de cuentas. En los tres casos, el énfasis varía: la primera línea es descriptiva, la segunda valorativa y la tercera normativa.

Finalmente, la influencia de Burrell (2016) genera una división entre autores que la reafirman y otros que la reformulan. Carabantes (2020) preserva la triple distinción

—intencional, técnica e interpretativa— como marco explicativo suficiente; Lo (2024) y Knoks & Raleigh (2022) la actualizan en el contexto de la IA profunda, pero advierten que las soluciones de explicabilidad (XAI) pueden producir una *falsa apariencia de comprensión* si no reproducen el razonamiento real del sistema. Estas tensiones reflejan el debate entre quienes buscan recuperar inteligibilidad y quienes priorizan la fiabilidad funcional de los modelos.

A partir de este análisis, proponemos la siguiente definición: La OA se entiende como la *ininteligibilidad práctica y normativa* de un sistema automatizado: la imposibilidad razonable de explicar o auditar sus decisiones con pruebas suficientes sobre entradas, reglas de inferencia y salidas. Dicha opacidad puede originarse por (i) limitaciones técnicas —complejidad, representaciones internas o uso de proxies—, (ii) déficits epistémicos —ausencia de métodos, herramientas o competencias—, (iii) dinámicas sociotécnicas —dependencia, pérdida de conocimiento o ocultamiento—, y (iv) restricciones institucionales y legales que afectan la trazabilidad y el significado de las operaciones.

Tal como se afirma en el párrafo inmediato anterior, la OA puede analizarse a partir de una taxonomía cuatridimensional que permite identificar con mayor precisión las fuentes y niveles en los que se manifiesta la ininteligibilidad de los sistemas automatizados. Esta taxonomía no busca sustituir las clasificaciones previas, sino integrarlas en un marco coherente que articule las causas técnicas, epistémicas, sociotécnicas e institucionales que dificultan la transparencia y la rendición de cuentas.

Las limitaciones técnicas constituyen el núcleo más visible de la opacidad. Incluyen la complejidad arquitectónica de los modelos, las representaciones internas opacas y el uso de proxies o variables sustitutivas. Mann et al. (2023) denominan a este plano opacidad arquitectónica, en la medida en que la densidad de capas y la interacción no lineal entre nodos dificultan la reconstrucción causal del proceso. Este fenómeno se agrava en los modelos de aprendizaje profundo, donde el número de parámetros supera la capacidad humana de rastreo. Carabantes (2020) y Lo (2024) advierten que tales estructuras emulan funciones cognitivas del cerebro sin articular razonamientos verificables, lo que las convierte en auténticas cajas negras incluso para sus desarrolladores. La opacidad técnica, por tanto, no depende de un ocultamiento deliberado, sino de la naturaleza misma del diseño algorítmico.

Los déficits epistémicos refieren a la ausencia o insuficiencia de métodos, herramientas y competencias que permitan comprender o justificar el funcionamiento de los sistemas. Freiman et al. (2025) diferencian entre opacidad epistémica —derivada de los límites cognitivos o metodológicos— y opacidad no epistémica —vinculada a la mala calidad de los datos o a normas defectuosas—. Petrolo et al. (2023) formalizan

esta dimensión mediante el modelo IPO (Input–Process–Output), que permite examinar la justificación epistémica en cada fase del algoritmo. No obstante, la falta de alfabetización digital e interdisciplinaria de muchos operadores jurídicos convierte esta dimensión en un obstáculo práctico para la transparencia institucional. En el ámbito judicial, ello se traduce en la dependencia de peritos o terceros técnicos, lo cual desplaza el control del juez hacia actores no sometidos al escrutinio jurisdiccional.

Las dinámicas sociotécnicas agrupan los procesos de dependencia, pérdida de conocimiento y ocultamiento que surgen de la interacción entre personas, instituciones y tecnologías. Mann et al. (2023) las definen como opacidad sociotécnica, resultado de la fragmentación del conocimiento entre diseñadores, usuarios e implementadores. Alvarado (2023) amplía este enfoque al denunciar que dichas dinámicas reproducen asimetrías cognitivas y exclusiones epistémicas, en tanto ciertos grupos quedan fuera de la comprensión y control de los sistemas. De igual modo, Gerdes (2021) distingue entre opacidad deliberada —intencionalmente impuesta por los desarrolladores— y opacidad emergente —resultado imprevisto del uso colectivo—. En ambos casos, la falta de trazabilidad de las decisiones tecnológicas erosiona la agencia humana y debilita los mecanismos de rendición de cuentas democráticos.

Finalmente, las restricciones institucionales y legales comprenden aquellas condiciones que impiden el acceso, la auditabilidad o la interpretación de los sistemas por razones de secreto comercial, propiedad intelectual o regulación insuficiente. Kroll (2018) identifica este plano como un problema de gobernanza algorítmica, donde la posibilidad de auditar depende tanto del diseño técnico como de los marcos normativos que lo regulan. Facchini y Termine (2022) lo expresan en su categoría de opacidad de acceso, mientras que Gorwa et al. (2020) lo desarrollan bajo las nociones de opacidad técnica, institucional y normativa. En contextos judiciales, tales restricciones pueden provenir de la adopción de software propietario sin mecanismos de supervisión pública, lo que contraviene los principios de transparencia y control jurisdiccional efectivo.

Esta taxonomía permite comprender que la OA no se agota en la ininteligibilidad técnica, sino que constituye un fenómeno multinivel y relacional, donde factores cognitivos, organizacionales y jurídicos se entrelazan. En conjunto, estas dimensiones explican por qué la simple exigencia de transparencia no basta: es necesario identificar en qué nivel se produce la falta de inteligibilidad para diseñar estrategias diferenciadas de mitigación —explicabilidad técnica, alfabetización epistémica, control sociotécnico y regulación institucional—.

e-ISSN 2382-4220 / Vol.28, n°. 3 / Especial DIXI -RI/INS 2026
Universidad Cooperativa de Colombia

Aceptabilidad vs. Explicabilidad de la OA.

A partir de los resultados de la Tabla 1 se identifica también una controversia en torno a las exigencias de explicabilidad de la IA versus una eventual aceptación de la OA. Kumar et al. (2024) introducen de manera explícita la noción de *opacidad aceptable*. Esta idea reconoce que ciertos niveles de ininteligibilidad pueden tolerarse cuando concurren tres condiciones simultáneas: (i) inteligibilidad técnica razonable, (ii) fiabilidad verificable mediante auditorías externas, y (iii) control institucional efectivo sobre el sistema. Desde la perspectiva de dichos autores, la transparencia absoluta no constituye un requisito normativo universal, y la legitimidad puede sostenerse aun cuando el funcionamiento interno del algoritmo no resulte plenamente accesible, siempre que existan salvaguardas externas robustas.

Este planteamiento contrasta directamente con la postura de Bjerring et al. (2025), quienes rechazan que pueda hablarse de una opacidad aceptable en contextos donde la legitimidad democrática depende de la inteligibilidad pública. Para estos autores, la opacidad priva a los ciudadanos de la capacidad de someter a escrutinio racional las decisiones automatizadas, con lo cual afecta el núcleo deliberativo de la democracia. Así, mientras Kumar et al. (2024) admiten un equilibrio pragmático entre eficacia y control, Bjerring et al. (2025) sostienen que cualquier opacidad que obstaculice la justificación pública es incompatible con la legitimidad normativa. El contraste entre ambas posturas tensiona el debate entre enfoques funcionalistas y enfoques deliberativos de la transparencia.

Desde una postura más ecléctica, Langer & König (2023) no pretenden definir estándares de aceptabilidad, pero sí muestran que la opacidad produce efectos diferenciados según el actor involucrado: usuarios, afectados, diseñadores, implementadores o reguladores. Desde esta lógica, la aceptabilidad no es absoluta, sino relacional: depende de quién enfrenta la opacidad, con qué capacidades epistémicas y bajo qué cargas institucionales. Su contribución permite matizar el debate entre Kumar et al. (2024) y Bjerring et al. (2025) al introducir una perspectiva multiactor que revela desigualdades previas en la distribución del conocimiento.

Por su parte, Lu (2022) adopta un enfoque de derechos que estrecha aún más el margen de tolerancia hacia la opacidad. Para este autor, la falta de inteligibilidad vulnera el derecho a la información y la autodeterminación informativa, especialmente cuando no existen vías efectivas de apelación o de control. En esta clave, la opacidad no es un parámetro negociable ni sujeto a balances funcionales: cualquier nivel que impida comprender los criterios relevantes de decisión resulta incompatible con la

protección de derechos fundamentales. Ello acerca la postura de Lu a la de Bjerring et al. (2025), aunque desde un fundamento constitucional antes que deliberativo.

A nivel ético, la postura de Hatherley (2025) también se opone a la idea de una opacidad aceptable cuando esta genera formas injustificadas de autoridad epistémica. Su planteamiento introduce la exigencia de justicia epistémica, según la cual la explicabilidad no solo informa, sino que redistribuye poder cognitivo entre quienes experimentan los efectos de la decisión automatizada. La explicación debe permitir disputar, cuestionar y revisar la autoridad del sistema. En ausencia de tales condiciones, la opacidad refuerza relaciones de dominación cognitiva.

Por su parte, Boge (2022) complejiza el debate al mostrar que incluso cuando un modelo es parcialmente explicable, puede no permitir *entender con él.* Esto plantea un límite estructural a la promesa de la explicabilidad: aun si se transparentan los mecanismos internos, el modelo puede seguir siendo epistémicamente insuficiente para comprender fenómenos reales. Este señalamiento no rechaza ni acepta la opacidad, pero revela que la explicabilidad técnica no siempre resuelve la ininteligibilidad cognitiva, lo que refuerza la necesidad de prudencia institucional.

El debate sobre la explicabilidad también los encontramos en el trabajo de Fleisher (2022), quien se ocupa de la XAI (Inteligencia Artificial Explicable, según sus siglas en Inglés). Según este autor, exigir que las explicaciones reproduzcan exactamente el algoritmo subyacente constituye un error conceptual: la fidelidad total no es un requisito de la explicación en ciencias ni en XAI, siempre que el modelo explicativo permita identificar patrones causales relevantes, generar comprensión práctica y habilitar acciones de intervención, revisión o impugnación.

Este enfoque matiza la noción de opacidad aceptable defendida por Kumar et al. (2024). Si la explicación funciona como un modelo idealizado, su valor no depende de replicar el sistema, sino de su capacidad para situar al usuario —incluido el sujeto de la decisión judicial— en una posición epistémica suficiente para comprender, evaluar críticamente o cuestionar el resultado. A la vez, ofrece un contrapunto a la tesis deliberativa de Bjerring et al. (2025) y a las exigencias de inteligibilidad pública o justicia epistémica planteadas por Lu (2022) y Hatherley (2025), al mostrar que la comprensión significativa no requiere transparencia exhaustiva, sino modelos explicativos ajustados al contexto y a las capacidades de los agentes involucrados.

No obstante, la integración de XAI como modelo idealizado no elimina los riesgos asociados a su uso en entornos judiciales. La legitimidad de estos modelos depende de que las idealizaciones representen de forma suficientemente fiable los patrones causales relevantes en los que se sustenta la decisión automatizada, algo aún no demostrado para los casos complejos que caracterizan la actividad jurisdiccional.

Por ello, la contribución de Fleisher refuerza el fundamento epistemológico de la moratoria propuesta: aunque la XAI puede ofrecer comprensión funcional en determinados dominios, la distancia entre sus idealizaciones y la complejidad normativa y probatoria del proceso judicial exige un enfoque precautorio, hasta verificar que estos mecanismos cumplen con los estándares de debido proceso, garantía jurisdiccional y control epistémico exigibles en un Estado constitucional de derecho.

El trabajo de Fraser et al. (2022) introduce una perspectiva peculiar: sostiene que las explicaciones de sistemas de IA no deben evaluarse en abstracto, sino en función del fin legal concreto que deben servir. Esta tesis desplaza la discusión desde la pregunta general por la inteligibilidad hacia la utilidad jurídica específica de las explicaciones en litigios civiles, particularmente en materia de responsabilidad extracontractual.

Desde esta perspectiva, la opacidad no se valora únicamente por su magnitud técnica, sino por su impacto en la capacidad del sistema para aportar evidencia relevante y utilizable para demostrar elementos del caso, como causalidad, defecto del producto o negligencia en el diseño. Es decir, Fraser et al. (2022) entienden que la opacidad solo puede considerarse aceptable cuando no interfiere con la posibilidad de satisfacer las cargas probatorias esenciales del proceso. Esto aproxima su posición a la línea pragmática de Kumar et al. (2024), aunque con un énfasis distinto: mientras Kumar introduce criterios institucionales de control y auditoría, Fraser et al. vinculan la aceptabilidad directamente con la función procesal que debe cumplir la explicación.

En cuanto a la explicabilidad, Fraser et al. (2022) argumentan que las explicaciones relevantes para el derecho civil deben ser parsimoniosas, orientadas a demostrar hechos esenciales y adaptadas al estándar probatorio correspondiente. En litigios de responsabilidad, no es necesario comprender todo el sistema, sino mostrar —con evidencia empírica y, cuando sea posible, con XAI local o global— que ciertos errores, patrones o pesos del modelo son causalmente significativos para el daño alegado. Bajo esta lógica, la explicabilidad no es un fin en sí mismo, sino un medio epistémico para reconstruir la cadena causal en contextos donde la IA participa en la producción del daño.

Este aporte amplía la discusión desarrollada hasta ahora en el manuscrito. En primer lugar, introduce un criterio normativo distinto para evaluar la opacidad aceptable: la pregunta no es solo si el sistema es comprensible en general, sino si permite al tribunal reconstruir el *qué ocurrió*. En segundo lugar, la propuesta de Fraser et al. (2022) defiende la relevancia de enfoques empírico-contrafactuales de XAI para suplir déficits de explicabilidad estructural, mostrando que incluso con acceso limitado (*descriptive access* o *query access*), es posible generar evidencia operativa.

En tercer lugar, este enfoque fortalece la tesis precautoria del artículo: si la explicabilidad es valiosa solo en la medida en que permite cumplir con las cargas probatorias, y si actualmente los sistemas opacos no proporcionan medios suficientes para satisfacerlas —menos aún en contextos judiciales donde se debe demostrar causalidad, razonabilidad o defecto—, entonces la opacidad no puede considerarse aceptable en ámbitos donde la IA pueda influir en el razonamiento jurisdiccional o en la determinación de hechos. La propuesta de Fraser et al., en consecuencia, aporta una base adicional para justificar que la adopción de herramientas algorítmicas en tribunales requiere un estándar reforzado de control epistémico.

De este modo, la XAI orientada a fines jurídicos no reemplaza los debates sobre legitimidad democrática, derechos fundamentales o justicia epistémica expuestos por otros autores, sino que los complementa al mostrar que, incluso desde un enfoque estrictamente procesal, la opacidad sigue siendo un obstáculo incompatible con las exigencias de prueba, causalidad y debido proceso propias del derecho civil y, por extensión, del proceso jurisdiccional en general.

La OA frente a la tutela jurisdiccional efectiva. Defensa de una moratoria de carácter precautorio.

Las condiciones actuales de la IA impiden su incorporación segura y legítima en el ámbito judicial. La cuestión no se limita a la constatación técnica de que los sistemas opacos dificultan el escrutinio humano, sino que compromete dimensiones epistémicas, institucionales y normativas que se direccionan a un punto común: sin un estándar validado de explicabilidad, el uso de IA en decisiones jurisdiccionales resulta incompatible con el derecho humano a una decisión justa.

La opacidad no es un defecto contingente que pueda corregirse con mejoras incrementales, sino un rasgo estructural. La opacidad emerge de la complejidad técnica, las asimetrías epistémicas y las dinámicas sociotécnicas que atraviesan el ciclo de vida de los modelos. Knoks & Raleigh (2022) y Lo (2014) han desarrollado la manera en que la explicabilidad puede generar una ilusión de transparencia: un relato ordenado que simula comprensión sin ofrecer acceso real a los factores causales relevantes. Boge (2022) añade que incluso las explicaciones consideradas *suficientes* pueden fallar en permitir que los usuarios entiendan el fenómeno subyacente, mientras que Fleisher ha demostrado que las representaciones post hoc funcionan más

e-ISSN 2382-4220 / Vol.28, n°. 3 / Especial DIXI -RI/INS 2026
Universidad Cooperativa de Colombia

como modelos idealizados que como descripciones fieles del proceso algorítmico, lo que limita su aptitud para servir como fundamento probatorio o justificativo.

En el ámbito procesal, Fraser et al. (2022) introducen un elemento esencial: una explicación algorítmica solo adquiere relevancia jurídica cuando permite reconstruir hechos, inferencias y vínculos causales indispensables para satisfacer una carga probatoria. Sin esa capacidad, la explicación no cumple la función jurídica que la legitimaría. Este enfoque se conecta con la exigencia de motivación, razonabilidad y capacidad de contradicción que estructura el proceso jurisdiccional. Una decisión judicial basada en sistemas que no permiten identificar qué información fue relevante, por qué lo fue y de qué manera contribuyó al resultado, deja a la persona afectada sin la posibilidad real de controvertir la decisión, impugnar sus fundamentos o demostrar su incorrección. En efecto, lo que está en juego no es únicamente la inteligibilidad técnica, sino la posibilidad misma de ejercer el derecho de defensa.

La Corte Interamericana de Derechos Humanos (2016), tal como se desprende del Caso Maldonado Ordoñez vs. Guatemala, exige que toda persona pueda comprender las razones que justifican la decisión que le afecta, contar con un recurso efectivo para impugnarla y verificar que el razonamiento del tribunal se mantenga dentro de marcos de racionalidad pública y controlable. La Corte Interamericana ha sido clara en afirmar que la motivación judicial constituye un componente esencial del derecho a una decisión justa, pues permite someter la actuación estatal a parámetros de transparencia, verificabilidad y rendición de cuentas. En ausencia de explicabilidad suficiente, estos elementos desaparecen: la persona queda sometida a una decisión cuyo proceso de producción es impenetrable, sin posibilidad de comprender o contestar su lógica interna.

Frente a este panorama, insistir en la incorporación de sistemas de IA en decisiones judiciales sin haber validado previamente un estándar de explicabilidad implica trasladar al ámbito jurisdiccional un riesgo epistémico que resulta inaceptable en un Estado constitucional de derecho. La propia literatura de gobernanza algorítmica reconoce que ningún actor aislado —sea la judicatura, la academia, los desarrolladores o los colegios profesionales— tiene la capacidad de verificar de manera integral los sistemas. Esto exige procesos de evaluación colectiva, capaces de articular capacidades técnicas, jurídicas y metodológicas para determinar si una explicación es jurídicamente apta y epistémicamente fiable. La ausencia de tales mecanismos justifica un enfoque precautorio.

Por ello, la moratoria no debe entenderse como un rechazo a la innovación, sino como una medida institucional necesaria para proteger las garantías propias de la función jurisdiccional. Es razonable sostener que el Estado se abstenga de incorporar

sistemas de IA en decisiones judiciales hasta que exista un estándar validado de explicabilidad, elaborado mediante mesas de trabajo que articulen a la administración de justicia, las universidades, los colegios profesionales y los desarrolladores. Solo cuando estas instancias logren definir qué tipo de explicaciones son suficientes, cómo deben evaluarse, qué parámetros deben cumplir y qué mecanismos de auditoría pueden aplicarse, será posible hablar de una utilización legítima y compatible con el derecho humano a una decisión justa.

El enfoque precautorio que se propone en este trabajo es una respuesta a la OA, que es una barrera estructural que impide el derecho humano a una decisión justa y a la tutela judicial efectiva. Sin la capacidad de auditar y comprender los mecanismos internos, las partes y el tribunal se ven impedidos de controvertir la decisión o de impugnar sus fundamentos, un obstáculo que implica trasladar un riesgo epistémico inaceptable al ámbito jurisdiccional.

Como se anticipó, el principio de precaución está orientado a problemas ambientales. La experiencia comparada construye algunas líneas base para identificar el umbral de riesgo e incertidumbre admisible. Para la Corte Constitucional colombiana, en Sentencia C-293 de 2002, se deben cumplir cinco requisitos al momento de realizar un análisis fáctico de aquellas circunstancias que pueden significar un daño ambiental, a saber: 1.- Que exista peligro de daño; 2.- Que éste sea grave e irreversible; 3.- Que exista un principio de certeza científica, así ésta no sea absoluta; 4.- Que la decisión que la autoridad esté encamina a impedir la degradación del medio ambiente; 5.- Que el acto en que se adopte la decisión sea motivado.

La sentencia T-080 de 2017 va más allá. En este proceso se decide la constitucionalidad de la aspersión aérea con glifosato y los efectos que puede tener para la salud. Aquí, vale la pena resaltar que el accionante en este caso fue Martín Narváez Gómez en calidad de Capitán del resguardo indígena Carijona de Puerto Nare (Guaviare), quien demandó los derechos fundamentales a la consulta previa, a la integridad étnica y cultural, a la libre determinación, a la salud en conexión con la vida y al medio ambiente sano. Ante la evidencia presentada en el expediente, la Corte concluyó que el glifosato es una sustancia que tiene la potencialidad de afectar la salud humana como probable agente cancerígeno y, también, de forma muy peligrosa, el medio ambiente, y con ello, precisó la función del principio de precaución, que se aplica cuando el riesgo o la magnitud del daño generado o que puede sobrevenir no es conocido con anticipación, porque no hay manera de establecer, a mediano o largo plazo, los efectos de una acción, lo cual generalmente ocurre por la falta de certeza científica absoluta acerca de las precisas consecuencias de un fenómeno, un producto o un proceso. En este expediente llama la atención la intervención del Ministerio de

e-ISSN 2382-4220 / Vol.28, n°. 3 / Especial DIXI -RI/INS 2026
Universidad Cooperativa de Colombia

Salud, quien, en el marco del principio de precaución, recomendó suspender el uso del glifosato en aspersiones aéreas de cultivos ilícitos en todo el país, habida cuenta de las pruebas presentadas por el comité de expertos de la OMS, que evidenciaba la correlación entre fumigaciones y la generación del linfoma no-Hodgkin.

Ahora bien, no toda decisión judicial o administrativa estaría sujeta al principio de precaución en tanto, se trata de decisiones voluntarias o que no comprometen significativamente derechos fundamentales o el debido proceso. Ejemplo de esto se encuentra en el Reglamento General de Protección de Datos Europea, en su artículo 22 indica que toda persona "tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzcan efectos jurídicos en él o le afecte significativamente de modo similar", salvo celebrar o ejecutar un contrato entre el interesado y la administración pública, consentimiento explícito del interesado, y que no contraventan normas especiales que lo prohíban (Lasbleiz & Milkes, 2021).

El Consejo Constitucional francés, en 2018, tomó una decisión que nos puede aproximar al principio precautelatorio, en tanto ha indicado que, debido a la complejidad y significado incierto, la adopción de decisiones individuales basadas únicamente en un algoritmo no cumplía las finalidades constitucionales de accesibilidad e inteligibilidad de la ley. Adicionalmente, resaltó el tribunal francés, decisiones individuales obre la base de un algoritmo "autodidacta" o de "aprendizaje automático" significaría desconocer los "factores determinantes que llevaron a tomar una decisión. Adicional a eso, el Consejo Constitucional francés identificó tres garantías, derivadas del Código de Relaciones entre el Público y la Administración (Code des relations entre le public et l`Administration) a saber: Comunicación del algoritmo y su funcionamiento, recursos idóneos de impugnación y prohibición de utilizar datos sensibles (Lasbleiz & Milkes, 2021). De esta manera, conforme a la citada decisión judicial, los algoritmos de auto-aprendizaje o machine learning no están habilitados para tomar decisiones sin ningún tipo de intervención humana, por lo que sólo quedarían habilitados los algoritmos deterministas, que se caracterizan porque el diseñador y/o autoridad determina las reglas correspondientes para su funcionamiento (Lasbleiz & Milkes, 2021, p. 337).

Al analizar este aspecto en la producción normativa de la Unión Europea (2024), en la AI Act se articula una respuesta a este problema. En dicho instrumento se señala que los sistemas de IA pueden circular fácilmente a través de las fronteras, y reconoce que ciertos Estados miembros ya han explorado la adopción de normas nacionales para garantizar que la IA sea confiable y segura, y se desarrolle de acuerdo con las obligaciones de derechos fundamentales. Sin embargo, la Comisión teme que estas normas nacionales divergentes fragmenten el mercado interno y disminuyan

la certeza jurídica para los operadores, por lo que la AI Act busca asegurar un nivel consistente y alto de protección en toda la Unión.

Además, dicha norma comunitaria establece que Los sistemas de IA de alto riesgo solo deben introducirse en el mercado de la Unión, ponerse en servicio o utilizarse si cumplen ciertos requisitos obligatorios. Dichos requisitos deben garantizar que los sistemas de IA de alto riesgo disponibles en la Unión, o cuyo resultado se utilice de otra forma en la Unión, no planteen riesgos inaceptables para intereses públicos importantes de la Unión reconocidos y protegidos por el Derecho de la Unión. La norma en comento establece que la administración de justicia y los procesos democráticos son ámbitos donde ciertos sistemas de IA deben clasificarse como de alto riesgo, debido a 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. Para abordar específicamente los riesgos de posibles sesgos, errores y opacidad, se considera apropiado calificar como sistemas de IA de alto riesgo a 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 clasificación de alto riesgo se extiende también a los sistemas de IA destinados a ser utilizados por organismos de resolución alternativa de litigios para esos fines, siempre que los resultados de dichos procedimientos produzcan efectos jurídicos para las partes. Dicho marco regulatorio es consistente con desarrollos doctrinarios como lo expuesto por Ahern (2025) respecto de la gobernanza anticipatoria, oportuna e iterativa; para la autora, si bien facilitar la innovación puede ser un objetivo central, no debe olvidarse la aplicación del principio de precaución al abordar la innovación emergente.

Conclusiones

La cuestión de la OA no constituye un problema meramente técnico o una dificultad pasajera que pueda resolverse mediante ajustes incrementales en los sistemas de IA. Por el contrario, la opacidad es un rasgo estructural que afecta a los modelos en todas sus fases: desde el diseño arquitectónico y la disponibilidad de métodos explicativos hasta las dinámicas sociotécnicas de implementación y las limitaciones institucionales que impiden su auditoría. Esta naturaleza estructural conduce a una primera conclusión: no existe, en la actualidad, un marco epistémico o regulatorio que garantice que los sistemas automatizados puedan satisfacer las exigencias de motivación, transparencia y control que caracterizan a una decisión jurisdiccional compatible con un Estado constitucional de derecho.

e-ISSN 2382-4220 / Vol.28, n°. 3 / Especial DIXI -RI/INS 2026
Universidad Cooperativa de Colombia

La postura funcionalista que propone la posibilidad de una opacidad aceptable, se muestra insuficiente frente a los estándares de inteligibilidad pública, justicia epistémica y a las exigencias de derechos fundamentales. Estas divergencias reflejan un debate vivo, pero coinciden en un punto esencial: no puede admitirse la delegación de funciones jurisdiccionales en sistemas cuyo funcionamiento no se encuentre disponible para el escrutinio racional, la contradicción efectiva y la verificación probatoria. La XAI, incluso en sus versiones más avanzadas, todavía genera modelos idealizados que no reproducen de manera fiel los mecanismos causales del sistema y, por ello, no garantizan el nivel de comprensión requerido para legitimar una decisión judicial.

Una explicación sólo es jurídicamente relevante si permite reconstruir hechos, vínculos causales y elementos esenciales para la satisfacción de las cargas probatorias. Mientras los sistemas opacos sigan sin ofrecer esta capacidad —y mientras no exista evidencia empírica de que pueden hacerlo en casos complejos—, su uso en decisiones judiciales equivale a incorporar un riesgo epistémico incompatible con el derecho a la defensa, la motivación y el control jurisdiccional. Esta línea argumental coincide con la jurisprudencia interamericana, particularmente con el Caso Maldonado Ordoñez vs. Guatemala, donde la Corte IDH afirmó que toda persona tiene derecho a comprender las razones que sustentan la decisión que le afecta, exigir un control judicial efectivo y verificar la racionalidad pública del razonamiento judicial.

Así, resulta razonable concluir que la incorporación de sistemas de IA en el ámbito judicial debe ser sometida a una moratoria de carácter precautorio. Esta moratoria no implica un rechazo a la innovación, sino una medida institucional destinada a garantizar que las decisiones jurisdiccionales continúen sustentándose en principios de transparencia, motivación y rendición de cuentas. La complejidad técnica de los modelos, unida a las limitaciones actuales de la XAI y a la ausencia de estándares validados de explicabilidad, impide asegurar que las decisiones automatizadas puedan cumplir con las garantías propias del debido proceso y la tutela judicial efectiva.

Finalmente, el camino hacia una eventual integración legítima de estas tecnologías requiere procesos deliberativos y de evaluación colectiva. Es imprescindible convocar mesas de trabajo que involucren a la administración de justicia, la academia, los colegios profesionales y los desarrolladores, con el fin de elaborar estándares de explicabilidad verificables, métodos de auditoría independientes y mecanismos institucionales de control. Mientras no se logre una validación sujeta a escrutinio público, la moratoria constituye no solo una medida prudente, sino una exigencia derivada del derecho humano a una decisión justa y del deber estatal de proteger la integridad del proceso judicial.

Referencias

Ahern, D. (2025). The New Anticipatory Governance Culture for Innovation: Regulatory Foresight, Regulatory Experimentation and Regulatory Learning. *European Business Organization Law Review*, (26), 241–283. https://doi.org/10.1007/s40804-025-00348-7

Alvarado, R. (2023). AI as an Epistemic Technology. *Science and Engineering Ethics, 29*(5), 32. https://doi.org/10.1007/s11948-023-00451-3

Bjerring, J. C., Mainz, J., & Munch, L. (2025). Deep learning models and the limits of explainable artificial intelligence. *Asian Journal of Philosophy, 4*(1), 1–26. https://doi.org/10.1007/s44204-024-00238-8

Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines, 32*(1), 43–75. https://doi.org/10.1007/s11023-021-09569-4

Cantarini, P. (2024). Algocracy, algorithmic institutionalism, digital rationality and risk to democracy. *Revista Juridica Unicuritiba, 4*(80), 590–626. https://doi.org/10.26668/revistajur.2316-753X.v4i80.7823

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & Society, 35*(2), 309–317. https://doi.org/10.1007/s00146-019-00888-w

Corte Interamericana de Derechos Humanos. (2016). *Caso Maldonado Ordoñez vs. Guatemala.* https://www.corteidh.or.cr/docs/casos/articulos/seriec_311_esp.pdf

Drnas de Clément, Z. (2008). *Elementos Esenciales del Principio de Precaución Ambiental.* In Universidad de Córdoba. Facultad de Derecho y Ciencias Sociales. Centro de Investigaciones Jurídicas y Sociales, *Anuario X 2007* (1st ed., pp. 329 - 340). Buenos Aires: Universidad de Córdoba. Facultad de Derecho y Ciencias Sociales. Centro de Investigaciones Jurídicas y Sociales.

Engstrom, D. F. (2024). The Automated State: A Realist View. *George Washington Law Review, 92*(6), 1437–1472.

Facchini, A., & Termine, A. (2022). Towards a Taxonomy for the Opacity of AI Systems. In *Studies in Applied Philosophy Epistemology and Rational Ethics* (Vol. 63, pp. 73–89). https://doi.org/10.1007/978-3-031-09153-7_7

Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme, 19*(4), 534–560. https://doi.org/10.1017/epi.2022.39

Fraser, H., Simcock, R., & Snoswell, A. J. (2022). AI Opacity and Explainability in Tort Litigation. *2022 ACM Conference on Fairness Accountability and Transparency,* 185–196. https://doi.org/10.1145/3531146.3533084

Freiman, O., McAndrews, J., Mansell, J., & van der Linden, C. (2025). 'Opacity' and 'Trust': From Concepts and Measurements to Public Policy. *Philosophy & Technology, 38*(1), 29. https://doi.org/10.1007/s13347-025-00862-z

Gerdes, A. (2021). Dialogical Guidelines Aided by Knowledge Acquisition: Enhancing the Design of Explainable Interfaces and Algorithmic Accuracy. In *Advances in Intelligent Systems and Computing* (Vol. 1288, pp. 243–257). https://doi.org/10.1007/978-3-030-63128-4_19

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945

Grant, D. G., Behrends, J., & Basl, J. (2025). What we owe to decision-subjects: beyond transparency and explanation in automated decision-making. *Philosophical Studies, 182*(1), 55–85. https://doi.org/10.1007/s11098-023-02013-6

Hatherley, J. (2025). A moving target in AI-assisted decision-making: dataset shift, model updating, and the problem of update opacity. *Ethics and Information Technology, 27*(2), 20. https://doi.org/10.1007/s10676-025-09829-2

Kaas, M. H. L. (2024). The perfect technological storm: artificial intelligence and moral complacency. *Ethics and Information Technology, 26*(3), 49. https://doi.org/10.1007/s10676-024-09788-0

Knoks, A., & Raleigh, T. (2022). XAI and philosophical work on explanation: A roadmap. *Ceur Workshop Proceedings, 3319*, 101–106. https://ceur-ws.org/Vol-3319/paper11.pdf

Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180084. https://doi.org/10.1098/rsta.2018.0084

Kumar, M., Aijaz, A., Chattar, O., Shukla, J., & Mutharaju, R. (2024). Opacity, Transparency, and the Ethics of Affective Computing. *IEEE Transactions on Affective Computing, 15*(1), 4–17. https://doi.org/10.1109/TAFFC.2023.3278230

Langer, M., & König, C. J. (2023). Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. *Human Resource Management Review, 33*(1), 100881. https://doi.org/10.1016/j.hrmr.2021.100881

Lasbleiz N. & Milkes I., (2021) Los desafíos de la automatización de las decisiones individuales por la Administración Pública, en "Disrupción tecnológica, transformación digital y sociedad, Ed. Externado.

Lee, F. (2025). The practices and politics of machine learning: a field guide for analyzing artificial intelligence. *AI & SOCIETY.* https://doi.org/10.1007/s00146-025-02430-7

Lo, F. T. H. (2024). The paradoxical transparency of opaque machine learning. *AI and Society, 39*(3), 1397–1409. https://doi.org/10.1007/s00146-022-01616-7

Lu, S. (2022). Data Privacy, Human Rights, and Algorithmic Opacity. *California Law Review, 110*(6), 2087–2147. https://doi.org/10.15779/Z38804XM07

Mann, S., Crook, B., Kastner, L., Schomacker, A., & Speith, T. (2023). Sources of Opacity in Computer Systems: Towards a Comprehensive Taxonomy. *Proceedings 31st IEEE International Requirements Engineering Conference Workshops Rew 2023,* 337–342. https://doi.org/10.1109/REW57809.2023.00063

Petrolo, M., Kubyshkina, E., & Primiero, G. (2023). A logical approach to algorithmic opacity. *Ceur Workshop Proceedings, 3615*, 89–95. https://ceur-ws.org/Vol-3615/short4.pdf

Riechmann, J & Tickner, J (2002). *El principio de precaución.* Barcelona: Icaria Editorial (pp. 47). Citado por González Villa, J (2006). *Derecho Ambiental Colombiano: Parte General Tomo I.* Bogotá D.C: Editorial Universidad Externado de Colombia.

Ruiz-Jarabo, D. (2005). El desarrollo comunitario del principio de precaución. En C. G. Judicial (Ed.). El principio de precaución y su proyección en el derecho administrativo español (pp. 41-74). Madrid: Centro de Documentación Judicial.

Unión Europea (2024). Reglamento de Inteligencia Artificial. Reglamento (Ue) 2024/1689 del Parlamento Europeo y del Consejo. https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=OJ:L_202401689