

Hacia una teoría jurídica de los fallos algorítmicos y su impacto diferenciado en los derechos humanos desde el PIDCP*

Towards a Legal Theory of Algorithmic Decision-Making and Its Differential Impact on Human Rights under the ICCPR

Rodolfo Moreno Cruz**

Resumen: El creciente uso de sistemas de inteligencia artificial en actividades relacionadas con la administración de justicia plantea nuevos desafíos para la protección de los derechos humanos. Sin embargo, gran parte de la literatura especializada analiza los riesgos algorítmicos de manera uniforme, sin distinguir la naturaleza específica de los distintos tipos de fallos que pueden producirse. El presente trabajo propone una clasificación jurídica de los fallos algorítmicos en tres categorías: errores técnicos, sesgos algorítmicos y alucinaciones. A partir de una metodología dogmático-jurídica e interdisciplinaria, se examinan las consecuencias diferenciadas que cada una de estas variantes puede generar sobre derechos protegidos por el Pacto Internacional de Derechos Civiles y Políticos. Finalmente, se cierra con una interrogante constitucional: cómo asegurar que, incluso en escenarios crecientemente automatizados, las personas continúen siendo tratadas como sujetos de derechos y no únicamente como datos procesables dentro de sistemas algorítmicos.

Abstract: The increasing deployment of artificial intelligence systems within activities related to the administration of justice poses novel challenges for the protection of human rights. However, much of the specialized literature analyzes algorithmic risks uniformly, failing to distinguish the specific nature of the distinct types of failures that may arise. This paper proposes a legal classification of algorithmic failures into three categories: technical errors, algorithmic biases, and hallucinations. Utilizing a dogmatic-legal and interdisciplinary methodology, this study examines the differentiated consequences that each of these variants may yield regarding the rights protected under the International Covenant on Civil and Political Rights. Finally, it concludes with a constitutional inquiry: how to ensure that, even in increasingly automated environments, individuals continue to be treated as subjects of rights and not merely as processable data within algorithmic systems

Palabras clave: Inteligencia artificial; fallos algorítmicos; Derechos humanos; Debido proceso; Pacto Internacional de Derechos Civiles y Políticos.

Keywords: Artificial Intelligence; algorithmic failures; Human Rights; Due Process International Covenant on Civil and Political Rights.

* Este trabajo se realizó en el marco del Proyecto «Estado de Derecho Algorítmico». Expreso mi agradecimiento a la Universidad Autónoma “Benito Juárez” de Oaxaca, así como a la Universidad de San Diego, la Universidad de Texas y la Universidad Carlos III de Madrid, por las facilidades brindadas para el acceso a sus bibliotecas durante el desarrollo de esta investigación. Agradezco también las valiosas observaciones y comentarios de mis estudiantes, colegas y de las dos personas revisoras anónimas, cuyas aportaciones contribuyeron a mejorar este trabajo. Cualquier error que permanezca en el texto es de mi exclusiva responsabilidad.

** Profesor del área de Posgrado en la Facultad de Derecho y Ciencias Sociales de la Universidad Autónoma Benito Juárez de Oaxaca (UABJO). Orcid: 0000-0003-3785-0044

Fecha de recepción: 01-06-2026

Fecha de aceptación: 31-07-2026

1. Introducción

La incorporación de herramientas de inteligencia artificial en actividades jurídicas y judiciales constituye una de las transformaciones más relevantes que han experimentado los sistemas de justicia contemporáneos. Algoritmos capaces de redactar documentos, analizar precedentes, clasificar riesgos o asistir procesos de decisión comienzan a ocupar espacios que históricamente pertenecían exclusivamente a operadores humanos del derecho. La promesa de eficiencia ha impulsado su expansión tanto en funciones administrativas como jurisdiccionales.

Sin embargo, junto con sus ventajas operativas, estas tecnologías también han introducido nuevos riesgos para los derechos humanos. Los sistemas de inteligencia artificial pueden producir resultados incorrectos. Cuando tales fallos impactan en procedimientos judiciales, las consecuencias dejan de ser exclusivamente tecnológicas y adquieren relevancia jurídica y convencional.

Existe escasa reflexión sistemática sobre cómo distintos tipos de fallos algorítmicos producen afectaciones diferenciadas a los derechos humanos y sobre cuáles son las consecuencias jurídicas específicas que derivan de cada uno de ellos.

Precisamente por ello, el presente ensayo parte de una distinción conceptual fundamental. No todos los fallos de la inteligencia artificial responden a la misma lógica ni generan idénticos riesgos jurídicos. Con fines analíticos, el trabajo distingue entre tres categorías principales: error técnico, sesgo algorítmico y alucinación¹. La relevancia de esta clasificación no es únicamente tecnológica. Cada una de estas formas de fallo compromete de manera distinta derechos humanos como la igualdad, el debido proceso, la

¹ Para efectos del presente ensayo, no se incluyen dentro de la propuesta de fallos los incidentes de manipulación externa mediante “prompt injection”, es decir, intentos deliberados de engañar al sistema a través de instrucciones maliciosas. Piénsese en el caso brasileño de las abogadas que intentaron inducir a una IA a producir respuestas contrarias a su diseño. Tampoco se consideran los errores humanos derivados de atribuir a las máquinas capacidades inexistentes (Ação Trabalhista Ordinária (ATOrd) nº 0001062-55.2025.5.08.0130, 3ª Vara do Trabalho de Parauapebas, 2026, p. 5). Un ejemplo paradigmático es el famoso caso mexicano del detector molecular GT-200, aparato al que se le adjudicaron cualidades científicas que nunca tuvo, y cuyo uso se basaba en una falsa expectativa tecnológica más que en un fallo algorítmico real (Facultad de Atracción 2/2012, Primera Sala de la Suprema Corte de Justicia de la Nación, 2012, p. 15). Estos episodios constituyen ataques o abusos del entorno humano contra el sistema, más que errores propios de la lógica interna de la inteligencia artificial. En consecuencia, la categoría de fallos que se analiza en este trabajo se centra exclusivamente en aquellos que emanan del funcionamiento interno de la IA.

presunción de inocencia y el acceso efectivo a la justicia. Asimismo, sus consecuencias pueden manifestarse tanto en decisiones jurisdiccionales como en actuaciones procesales realizadas por autoridades, abogados o particulares.

En este contexto, el ensayo sostiene que el Pacto Internacional de Derechos Civiles y Políticos (PIDCP)² ofrece un marco normativo robusto para enfrentar los riesgos derivados del uso de inteligencia artificial en el ámbito jurídico. Aunque dicho instrumento fue elaborado con anterioridad al desarrollo de los sistemas algorítmicos contemporáneos, sus principios conservan capacidad para proyectarse hacia nuevas formas de afectación tecnológica y para estructurar estándares de protección de derechos humanos.

A partir de estas premisas, el trabajo examina, en primer término, las principales categorías de fallos algorítmicos y sus características distintivas. Posteriormente, analiza el impacto jurídico específico de cada una de ellas sobre los derechos humanos reconocidos por el PIDCP. Finalmente, propone algunos criterios orientados a construir un modelo de responsabilidad jurídica compatible con las exigencias del Estado constitucional y convencional de derecho frente al avance de la inteligencia artificial.

2. Tipología conceptual de los fallos algorítmicos

La clasificación desarrollada en este trabajo constituye una herramienta conceptual para el análisis jurídico de la inteligencia artificial. Antes de examinar las consecuencias que cada tipo de fallo produce sobre los derechos humanos, resulta necesario delimitar las características del error técnico, el sesgo algorítmico y la alucinación.

2.1. El error técnico

El error técnico consiste, en términos generales, en una falta de precisión entre lo que el sistema predijo y lo que correspondía al resultado correcto. Si una herramienta diseñada para identificar imágenes clasifica a un gato como perro, existe un error. Se trata de una desviación objetivamente medible³ y, en muchos casos, corregible mediante mejores datos, ajustes del modelo o entrenamiento adicional⁴. Por ejemplo, unos investigadores diseñaron una figura tridimensional con forma de tortuga y patrones visuales específicos. Para cualquier observador humano era claramente una tortuga. Sin embargo, diversos modelos de

² Pacto Internacional de Derechos Civiles y Políticos, 1966.

³ Bishop, C.M., *Pattern Recognition and Machine Learning*, 1a ed., Nueva York: Springer, 2006, pp. 4-20.

⁴ Géron, A., *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2.^a ed., Sebastopol (CA): O'Reilly Media, 2019, pp. 82-83.

clasificación de imágenes la identificaban como rifle⁵. El episodio demostró cómo algunos sistemas pueden reconocer patrones matemáticos sin comprender verdaderamente el objeto que tienen enfrente.

Este tipo de errores no se limita al ámbito experimental. En la práctica, sistemas de inteligencia artificial han mostrado fallas en tareas aparentemente simples: desde errores en operaciones aritméticas básicas hasta problemas de reconocimiento de voz. Se trata de desviaciones técnicas. En este caso (de error técnico) el sistema no inventa información, simplemente se equivoca.

Ahora bien, el problema adquiere una dimensión distinta cuando el error ocurre en contextos jurídicos o administrativos. Una equivocación aparentemente menor puede alterar declaraciones judiciales, cálculos procesales o valoraciones probatorias con consecuencias relevantes para los derechos de las personas. En estos escenarios, la precisión deja de ser únicamente una cuestión tecnológica y se convierte en una exigencia jurídica vinculada al debido proceso y a la seguridad jurídica.

2.2. El sesgo algorítmico

Si el error técnico responde esencialmente a un problema de precisión, el sesgo algorítmico plantea una dificultad distinta y más profunda. Aquí no estamos frente a una equivocación aislada o azarosa, sino ante una tendencia sistemática que perjudica de manera desproporcionada a determinados sectores sociales.

La noción contemporánea de sesgo posee importantes antecedentes en la psicología cognitiva y en los estudios sobre racionalidad limitada desarrollados por Daniel Kahneman y Amos Tversky⁶. Posteriormente, el concepto fue trasladado al ámbito estadístico y computacional para describir desviaciones sistemáticas dentro de modelos predictivos y sistemas automatizados de decisión.

En inteligencia artificial, el sesgo suele originarse porque los sistemas aprenden a partir de datos históricos producidos por sociedades que ya contienen desigualdades estructurales. Como consecuencia, los algoritmos pueden reproducir —e incluso amplificar— prejuicios previamente existentes bajo una apariencia de neutralidad matemática.

⁵ Athalye, A., Engstrom, L., Ilyas, A. y Kwok, K., “Synthesizing Robust Adversarial Examples”, en J. Dy y A. Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 284–285.

⁶ Kahneman, D. y Tversky, A., “Prospect Theory: An Analysis of Decision under Risk”, *Econometrica*, vol. 47, nº 2, 1979, pp. 284–293.

En términos informáticos, el sesgo suele originarse en los datos de entrenamiento, o en la forma en que se definen los objetivos⁷. Si una base de datos contiene una representación mínima de personas adultas mayores o comunidades indígenas, el sistema aprenderá menos sobre esos grupos y tenderá a tratarlos peor. De igual forma, si un modelo es optimizado solo para “máxima eficiencia” sin incorporar criterios de equidad, puede generar resultados aparentemente exitosos, pero socialmente discriminatorios.

Un ejemplo sencillo puede ayudar a comprenderlo. Imaginemos una aplicación escolar que recomienda becas deportivas con base en datos históricos. Si durante años la mayoría de las becas se otorgaron a hombres porque casi no se impulsaba el deporte femenino, la herramienta podría concluir que los varones “tienen mayor probabilidad de éxito” y comenzar a recomendar casi exclusivamente a estudiantes hombres. El sistema parece eficiente, pero en realidad está repitiendo una desigualdad pasada. No se trata de un error casual: es un sesgo aprendido.

Precisamente ahí radica la diferencia central entre error técnico y sesgo algorítmico. Mientras el primero se refiere principalmente a inexactitud, el segundo obliga a preguntarse cómo se distribuyen los errores y quiénes soportan sus consecuencias con mayor intensidad.

Esta distinción resulta especialmente relevante desde la perspectiva de los derechos humanos. Un sistema puede mostrar niveles generales aceptables de precisión estadística y, al mismo tiempo, producir afectaciones desproporcionadas sobre grupos históricamente vulnerables. En tales escenarios, el problema deja de ser exclusivamente tecnológico y entra directamente en el terreno de la igualdad y la no discriminación.

2.3. La alucinación

La alucinación constituye uno de los fenómenos más característicos de los modelos contemporáneos de lenguaje generativo. A diferencia del error técnico y del sesgo algorítmico, aquí el problema no consiste únicamente en producir resultados incorrectos o discriminatorios, sino en fabricar información inexistente con apariencia de autenticidad.

La alucinación ocurre cuando un sistema genera datos, precedentes, hechos o afirmaciones falsas presentadas con coherencia lingüística y apariencia de autoridad⁸. El fenómeno deriva de la propia arquitectura probabilística de

⁷ Barocas, S. y Selbst, A.D., “Big Data’s Disparate Impact”, *California Law Review*, vol. 104, n° 3, 2016, pp. 680–683.

⁸ Alansari, A. y Luqman, H., “Large language models hallucination: A comprehensive survey”, *Computer Science Review*, vol. 61, Ámsterdam: Elsevier, 2026, art. 100970, pp. 1–2.

muchos modelos de lenguaje, cuyo objetivo principal no es verificar la verdad de una afirmación, sino producir la secuencia lingüística estadísticamente más probable.

Precisamente por ello, la alucinación puede resultar especialmente peligrosa en contextos jurídicos. Un error evidente suele despertar sospechas inmediatas; pero no las levanta una afirmación falsa redactada con precisión técnica y seguridad argumentativa.

Un ejemplo paradigmático ocurrió en el caso *Mata v. Avianca, Inc.*, donde abogados utilizaron ChatGPT para localizar precedentes judiciales aplicables a una demanda civil en Estados Unidos⁹. El sistema proporcionó casos inexistentes, con citas, números de expediente y razonamientos completamente ficticios. Lo particularmente problemático fue que la herramienta insistió reiteradamente en la autenticidad de dichos precedentes a pesar de que los abogados le solicitaron confirmación adicional.

El episodio evidenció que los sistemas contemporáneos de inteligencia artificial no solo pueden equivocarse, sino también producir información jurídicamente falsa con apariencia plenamente legítima. En consecuencia, la alucinación plantea riesgos específicos para la verdad procesal¹⁰.

3. Impacto jurídico de los fallos algorítmicos

El problema del error dentro del derecho no es nuevo. Desde el derecho romano ya existía preocupación por los efectos jurídicos del error y de los actos viciados. Concretamente el Digesto de Justiniano recogía la máxima *quod initio vitiosum est, non potest tractu temporis convalescere* (“Lo que está viciado desde su inicio no puede convalidarse con el transcurso del tiempo”). Aunque formulada siglos antes del desarrollo tecnológico contemporáneo, esta preocupación conserva vigencia frente a sistemas algorítmicos que pueden introducir fallos estructurales dentro de procedimientos jurídicos aparentemente válidos.

Ahora bien, el error acompaña a los seres humanos de forma ineludible pero históricamente ha existido una especial preocupación por el error judicial, es decir, por aquel que se materializa en la decisión final adoptada por quien tiene

⁹ *Mata v. Avianca, Inc.*, United States District Court for the Southern District of New York, 22 de junio de 2023, caso n° 1:2022cv01461 (PKC).

¹⁰ Es decir, en palabras de Tuzet “lo que interesa en el ámbito del proceso y de la prueba jurídica son las pretensiones fácticas susceptibles de ser empíricamente verificadas o falsadas” en Tuzet, G., *Filosofía de la prueba jurídica*, 1.ª ed., Madrid: Marcial Pons, 2021, p. 85. Sobre este punto también puede consultarse a Ferrer Beltrán, J., *La valoración racional de la prueba*, Madrid: Marcial Pons, 2005, p. 62.

la autoridad de resolver el conflicto¹¹. La razón es evidente: el error del juzgador posee capacidad directa para afectar libertad, patrimonio, igualdad o incluso, parafraseando a Moreso, define la salud de un sistema jurídico¹².

Precisamente por ello, el derecho internacional de los derechos humanos nunca ha partido de la premisa irreal según la cual los sistemas de justicia son infalibles. El propio sistema de Naciones Unidas, a través del Pacto Internacional de Derechos Civiles y Políticos, reconoce implícitamente la posibilidad del error judicial. El Comité de Derechos Humanos ha sostenido expresamente que el artículo 14:

“garantiza únicamente la igualdad y la imparcialidad en los procedimientos judiciales y no puede ser interpretado en el sentido de que garantiza la ausencia de errores de parte del tribunal competente”(ONU, 2007, p. 10)¹³.

No obstante, el reconocimiento de la inevitabilidad del error no implica indiferencia frente a sus consecuencias. Por el contrario, el propio PIDCP establece, en su artículo 14.6, el derecho a indemnización cuando una persona ha sido condenada mediante sentencia firme y posteriormente se demuestra la existencia de un error judicial.

Por ello, la cuestión central no consiste en determinar si los sistemas algorítmicos pueden equivocarse —pues el error forma parte inevitable de toda actividad humana y tecnológica—, sino en comprender qué tipo de errores producen, cuáles son las consecuencias jurídicas diferenciadas que generan dentro de los procedimientos judiciales.

En este punto, la discusión contemporánea adquiere una dimensión particularmente compleja. Como ha señalado Ahti Saarenpää, el funcionamiento del Estado constitucional contemporáneo depende profundamente de sistemas confiables de comunicación e información jurídica¹⁴. El conocimiento jurídico no surge de manera espontánea, sino de procesos institucionales de validación de información. En consecuencia, cuando herramientas algorítmicas alteran información utilizada dentro del proceso judicial, el problema deja de ser únicamente tecnológico y pasa a comprometer directamente las condiciones de confiabilidad sobre las cuales descansa el propio Estado de derecho.

¹¹ Carbonell Bellolio, F., Meroi, A. y Uscanga, A. (eds.), *El error judicial. Problemas y regulaciones*, Valencia: Tirant lo Blanch, 2023, pp. 31–33.

¹² Carbonell Bellolio, F. y Fernández Blanco, C. (eds.), *Errores en la creación y aplicación del Derecho*, Madrid: Marcial Pons, 2025, p. 234.

¹³ ONU, *Observación general N° 32: Artículo 14. El derecho a un juicio imparcial y a la igualdad ante los tribunales y cortes de justicia*, Comité de Derechos Humanos, 2007, p. 10.

¹⁴ Saarenpää, A., “Law as Communication in a Digital Constitutional State”, en *Human-Machine Cooperation in Cyberspace*, Weblaw, 2026, pp. 29–39

Precisamente por ello, la clasificación de los fallos algorítmicos adquiere verdadera relevancia cuando se traslada al ámbito jurídico. No todos los fallos de inteligencia artificial afectan de la misma manera los procedimientos judiciales, ni todos comprometen idénticos derechos fundamentales.

En consecuencia, el impacto jurídico de la inteligencia artificial debe analizarse de manera diferenciada, atendiendo a la naturaleza específica del fallo algorítmico involucrado y a los derechos fundamentales potencialmente afectados. Pongamos por caso que cuando se producen errores técnicos, el principal riesgo recae sobre el debido proceso. Cuando aparecen sesgos algorítmicos, la afectación central se relaciona con el derecho a la igualdad y la prohibición de discriminación estructural. Finalmente, cuando los sistemas generan alucinaciones —esto es, información inexistente presentada con apariencia de autenticidad—, se comprometen garantías vinculadas a la tutela judicial efectiva.

A partir de esta diferencia, resulta posible construir un análisis jurídico más preciso sobre los riesgos contemporáneos de la inteligencia artificial en la administración de justicia y sobre las obligaciones estatales derivadas del derecho internacional de los derechos humanos frente a cada una de estas variantes de fallo algorítmico.

3.1. Interludio: bases convencionales para una teoría del error en derechos humanos

Existe abundante literatura sobre los errores judiciales y sus efectos en la administración de justicia, pero no hay un desarrollo explícito de una teoría convencional del error dentro del PIDCP. En gran medida, el tratado presupone la relevancia jurídica del error sin construir una teoría sistemática sobre él. Sin embargo, diversas disposiciones del propio Pacto, así como su interpretación por órganos internacionales y tribunales regionales, permiten identificar ciertos principios estructurales que sirven de base para identificar las bases convencionales de una teoría del error en materia de derechos humanos. Son 7 pilares.

Téngase presente lo siguiente: aunque el PIDCP no formula expresamente una teoría general del error judicial, ello no impide reconstruir sus presupuestos normativos mediante una interpretación sistemática e integradora de sus disposiciones. El ejercicio que aquí se propone no consiste en atribuir al Pacto contenidos inexistentes, sino en identificar, a partir de sus normas, de la Observación General núm. 32 del Comité de Derechos Humanos y de la interpretación desarrollada por la jurisprudencia internacional, los principios estructurales implícitos en el tratamiento convencional del error. Se trata, por tanto, de una reconstrucción doctrinal de carácter hermenéutico, cuyo propósito es sistematizar elementos dispersos del régimen convencional para ofrecer un marco analítico que permita comprender la relevancia jurídica del error en la protección internacional de los derechos humanos.

La selección de los siete pilares responde a un criterio funcional. Cada uno de ellos representa una dimensión distinta de la manera en que el sistema convencional previene, corrige, repara o limita las consecuencias jurídicas del error. En conjunto, estos elementos no pretenden describir una teoría expresamente formulada por el PIDCP, sino en reconstruir la arquitectura normativa subyacente a sus garantías procesales.

Primero. El derecho a no ser juzgado mediante procedimientos influenciados por error deriva del requisito de imparcialidad previsto en el artículo 14 del PIDCP. Así lo dispone la Observación General No. 32, párrafo 21, cuando señala: “El requisito de imparcialidad tiene dos aspectos. En primer lugar, los jueces no deben permitir que su fallo esté influenciado por sesgos o prejuicios personales, ni tener ideas preconcebidas en cuanto al asunto sometido a su estudio, ni actuar de manera que indebidamente promueva los intereses de una de las partes en detrimento de los de la otra”.

Segundo. El error puede ser inevitable, pero el derecho exige reducir razonablemente su probabilidad mediante mecanismos adecuados de transparencia, revisión, y control. La inevitabilidad del error no elimina el deber estatal de adoptar las mayores medidas posibles para prevenir afectaciones indebidas a derechos fundamentales. Su fundamento se encuentra en la Observación General No. 32 del Comité de Derechos Humanos, párrafo 26, cuya cita textual establece: “El artículo 14 garantiza únicamente la igualdad y la imparcialidad en los procedimientos judiciales y no puede ser interpretado en el sentido de que garantice la ausencia de errores por parte del tribunal competente”.

Tercero. Toda resolución dictada con base en un error judicial que produzca una condena indebida da lugar a una obligación de indemnización por parte del Estado. Su fundamento se encuentra en el artículo 14, numeral 6, del PIDCP: “Cuando una sentencia condenatoria firme haya sido ulteriormente revocada, o el condenado haya sido indultado por haberse producido o descubierto un hecho plenamente probatorio de la comisión de un error judicial, la persona que haya sufrido una pena como resultado de tal sentencia deberá ser indemnizada, conforme a la ley”.

Cuarto. La relevancia jurídica del error deriva de su capacidad para producir privaciones erróneas de derechos. El problema central no es únicamente la existencia abstracta de una equivocación técnica o humana, sino la posibilidad de que dicha equivocación afecte ilegítimamente libertad, patrimonio, acceso a la justicia, presunción de inocencia o incluso la vida de las personas.

Es cierto que esta idea no aparece formulada de manera expresa en el texto del PIDCP. El Pacto reconoce el error judicial — por ejemplo, en su artículo 14.6 — y establece diversas garantías orientadas a limitar actuaciones arbitrarias, pero

no desarrolla explícitamente una teoría general sobre las “privaciones erróneas de derechos”. Sin embargo, una interpretación sistemática del artículo 14 permite advertir que buena parte de las garantías procesales previstas en el Pacto — defensa adecuada, contradicción, revisión judicial, imparcialidad del tribunal y derecho a recurrir — sólo adquieren pleno sentido si se reconoce que lo que se busca es evitar privaciones erróneas de derechos.

Diversas tradiciones jurídicas han intentado explicar esta misma preocupación mediante categorías como juicio justo, tutela judicial efectiva, debido proceso o administración adecuada de justicia. No obstante, una de las formulaciones más influyentes y conceptualmente más precisas proviene de *Mathews v. Eldridge*, donde la Suprema Corte sostuvo que la suficiencia constitucional de un procedimiento depende, entre otros factores, del “*risk of an erroneous deprivation*” derivado de los mecanismos utilizados y del probable valor de salvaguardas adicionales destinadas a reducir dicho riesgo¹⁵.

Bajo esta lógica, las garantías procesales funcionan como mecanismos institucionales orientados a disminuir el riesgo de decisiones incorrectas antes de que estas produzcan consecuencias irreversibles sobre los derechos de las personas.

Ahora bien, la preocupación por las privaciones erróneas de derechos conduce inevitablemente a una cuestión propiamente iusfilosófica: ¿de qué depende la legitimidad de una decisión estatal que afecta derechos fundamentales? La transformación tecnológica de la función jurisdiccional obliga, en palabras de José Ignacio Solar Cayón, a imaginar un nuevo derecho¹⁶ capaz de responder a los desafíos de la sociedad digital. Desde distintas perspectivas, la filosofía del derecho ha intentado responder a esta cuestión. Para Jürgen Habermas, la legitimidad del derecho descansa en la posibilidad de que sus destinatarios participen, en condiciones de racionalidad, en procesos de deliberación y justificación pública¹⁷ o dicho en palabras del propio Habermas:

“La teoría discursiva del derecho explica la legitimidad del derecho con ayuda de procedimientos y presupuestos comunicativos –institucionalizados a su vez jurídicamente– que fundan la presunción de que los procesos de producción del derecho y aplicación del derecho conducen a resultados racionales”¹⁸

Por su parte, para Luigi Ferrajoli, la legitimidad del Estado constitucional busca “alcanzar valores éticos políticos con fines universales para la armonía de

¹⁵ *Mathews v. Eldridge*, 424 U.S. 319 (1976).

¹⁶ Solar Cayón, J.I., *Inteligencia artificial jurídica e imperio de la Ley*, Valencia: Tirant lo Blanch, 2025, p. 82.

¹⁷ Moreno Cruz, R., “Jürgen Habermas. Baedeker de su propuesta jurídica”, *Problema. Anuario de Filosofía y Teoría del Derecho*, nº 17, 2023, pp. 236–240.

¹⁸ Habermas, J., *Facticidad y validez*, 4.ª ed., Madrid: Trotta, 2005, p. 497.

la sociedad”¹⁹, escribe el autor italiano lo siguiente: “Por legitimación externa entiendo... criterios de valoración morales o políticos o de utilidad de tipo extra o meta-jurídico”²⁰.

Estas aproximaciones muestran que la legitimidad jurídica nunca ha dependido exclusivamente de la autoridad que decide, sino también de las condiciones que hacen racionalmente aceptable la decisión.

La irrupción de la inteligencia artificial introduce, sin embargo, una dimensión adicional a este debate. Como ha advertido Luciano Floridi, los sistemas algorítmicos pueden producir resultados altamente sofisticados sin “conciencia, intención o vida mental”,²¹ lo que obliga a preguntarse si una decisión construida mediante correlaciones estadísticas puede satisfacer las exigencias de justificación propias del razonamiento jurídico. Desde otra perspectiva, María José Belloso Martín ha mostrado que el empleo de algoritmos en la administración de justicia plantea interrogantes acerca de la reproducción de sesgos estructurales²² y de la posibilidad o imposibilidad de hablar de algoritmos justos²³. En este contexto, la propuesta de un humanismo tecnológico, defendida por Fernando H. Llano Alonso, ofrece un criterio especialmente valioso al referirse a lo que él denomina como “la transfiguración de la justicia humana en el universo Digital”²⁴.

En este contexto, el problema del error deja de ser una cuestión meramente técnica. La pregunta relevante ya no es solo si el sistema se equivocó, sino si el procedimiento que generó la decisión ofrecía garantías suficientes de legitimidad. Precisamente por ello, el análisis del riesgo de privación errónea adquiere una dimensión más profunda: las garantías procesales funcionan no solo como mecanismos de control formal, sino como instrumentos destinados a asegurar decisiones construidas mediante procedimientos racionalmente confiables y susceptibles de revisión crítica.

¹⁹ Moreno Cruz, R., “El modelo garantista de Luigi Ferrajoli: Lineamientos generales”, *Boletín Mexicano de Derecho Comparado*, vol. 40, n° 120, 2007, p. 842.

²⁰ Ferrajoli, L., *Derecho y Razón. Teoría del garantismo penal*, 5.ª ed., Madrid: Trotta, 2001, p. 213.

²¹ Floridi, L., “AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis”, *Philosophy & Technology*, vol. 38, n° 1, 2025, p. 30

²² Belloso Martín, N., “La problemática de los sesgos algorítmicos (con especial referencia a los de género). ¿Hacia un derecho a la protección contra los sesgos?”, en Llano Alonso, F. H. (dir.), Garrido Martín, J. y Valdivia Giménez, R. (coords.), *Inteligencia Artificial y Filosofía del Derecho*, Murcia: Editorial Laborum S.L., 2022, p. 46.

²³ Belloso Martín, N., “Sobre fairness y machine learning: el algoritmo ¿puede (y debe) ser justo?”, *Anales de la Cátedra Francisco Suárez*, núm. 57, 2023, pp. 7-38

²⁴ Llano Alonso, F., *Homo Ex Machina. Ética de la inteligencia artificial y Derecho digital ante el horizonte de la singularidad tecnológica*, Valencia: Tirant lo Blanch, 2024, p. 92.

Quinto. El derecho a recurrir constituye uno de los principales mecanismos institucionales de corrección del error. Desde esta perspectiva, la doble instancia y los mecanismos de revisión no representan únicamente garantías procesales autónomas, sino instrumentos orientados a detectar y corregir decisiones potencialmente incorrectas antes de que produzcan consecuencias irreversibles. Su fundamento puede encontrarse en el artículo 2, numeral 3, inciso a), del PIDCP, relativo al derecho a un recurso efectivo.

Sexto. El problema del error no se limita exclusivamente a condenas injustas. También las resoluciones erróneas de no responsabilidad o las decisiones sustentadas en valoraciones defectuosas pueden generar impunidad, y debilitar la confianza pública en el sistema de justicia. Una teoría integral del error no se limita únicamente a la protección frente a condenas indebidas; también debe considerar las consecuencias institucionales y sociales derivadas de decisiones incorrectas que impiden la determinación adecuada de responsabilidades. Sobre este aspecto el Conjunto de principios para la protección y la promoción de los derechos humanos mediante la lucha contra la impunidad define a la impunidad de la siguiente manera:

“Por impunidad se entiende la inexistencia, de hecho o de derecho, de responsabilidad penal por parte de los autores de violaciones, así como de responsabilidad civil, administrativa o disciplinaria, porque escapan a toda investigación con miras a su inculpación, detención, procesamiento y, en caso de ser reconocidos culpables, condena a penas apropiadas, incluso a la indemnización del daño causado a sus víctimas”.

Séptimo. Corresponde primariamente a los Estados parte revisar y corregir sus propios errores judiciales, salvo cuando estos resulten manifiestamente arbitrarios o equivalgan a una denegación de justicia. Su fundamento se encuentra en la Observación General No. 32, párrafo 26, que establece:

“En general, incumbe a los tribunales de los Estados Partes en el Pacto examinar los hechos y las pruebas o la aplicación de la legislación interna en cada caso particular, a menos que se demuestre que la evaluación de las pruebas o la aplicación de la legislación fue claramente arbitraria o equivalió a un error manifiesto o una denegación de justicia o que el tribunal incumplió de algún otro modo su obligación de independencia e imparcialidad”.

3.2. El reconocimiento jurídico de los riesgos algorítmicos

Los sistemas de inteligencia artificial no eliminan el problema histórico del error en la toma de decisiones; más bien trasladan ese riesgo al ámbito algorítmico y, en ocasiones, lo amplifican mediante la automatización y la opacidad técnica.

Por ello, distintos organismos nacionales e internacionales han buscado establecer marcos regulatorios que reconozcan y contengan estos riesgos. Un ejemplo paradigmático es el Reglamento de la IA de la Unión Europea²⁵ que en su considerando 61 clasifica como de alto riesgo aquellos sistemas de IA utilizados en el ámbito judicial o en mecanismos de resolución alternativa de litigios, precisamente por la posibilidad de sesgos, errores y falta de transparencia. El texto enfatiza que la IA puede apoyar la labor judicial, pero nunca sustituir la independencia ni la decisión final humana. Asimismo, en el anexo III: Sistemas de IA de alto riesgo contemplados en el apartado 2 del artículo 6 se enfatiza en este tema al referirse a la administración de justicia y procesos democráticos

En la misma línea, las Directrices de la UNESCO para el uso de sistemas de IA en cortes y tribunales recomiendan implementar un “análisis de riesgos algorítmicos”²⁶ para anticipar efectos positivos y negativos en la administración de justicia. Estas disposiciones reflejan la preocupación común por los errores algorítmicos y la situación de que no son meramente problemas técnicos, sino que pueden tener consecuencias jurídicas directas sobre los derechos de las personas.

La propuesta desarrollada en este trabajo parte de un diálogo multinivel entre fuentes normativas. El PIDCP aporta el fundamento convencional de las garantías procesales, mientras que el Reglamento Europeo de Inteligencia Artificial y las Directrices de la UNESCO concretan las obligaciones de prevención, supervisión y gestión de riesgos exigibles cuando dichas garantías deben proyectarse sobre sistemas algorítmicos.

Por ello resulta necesario distinguir entre los distintos tipos de fallos.

3.3. Errores técnicos y garantía del debido proceso

Los recientes errores técnicos de la IA han caminado desde lo cómico hasta lo peligroso. Por ejemplo, en 2024 Un cliente pidió un helado de caramelo y la máquina, con la obstinación de un chef experimental, insistió en añadir mantequilla. El resultado no fue un postre sino un performance gastronómico. Por otra parte, investigaciones relacionadas con vehículos autónomos documentaron casos de accidentes por errores del sistema²⁷.

²⁵ Unión Europea, Ley de Inteligencia Artificial (AI Act), 2024.

²⁶ UNESCO, *Directrices para el uso de sistemas de IA en Cortes y Tribunales*, 1.ª ed., París: UNESCO, 2026, p. 23.

²⁷ National Transportation Safety Board, *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018* (Highway Accident Report NTSB/HAR-19/03, PB2019-101402), Washington, D.C.: National Transportation Safety Board, 2019, p. 1

En contextos judiciales y administrativos, incluso desviaciones técnicas aparentemente menores pueden comprometer derechos fundamentales. Una transcripción automatizada incorrecta, un cálculo procesal defectuoso o un sistema erróneo de identificación pueden alterar decisiones con impacto directo sobre la libertad personal o el acceso efectivo a la justicia.

Pensemos, por ejemplo, en una herramienta automatizada utilizada para transcribir audiencias judiciales. Si el sistema convierte la expresión “no acepto los hechos” en “acepto los hechos”, el problema deja de ser una simple imprecisión tecnológica. La alteración sustancial del sentido de una declaración puede comprometer el derecho de defensa, y modificar el resultado mismo del procedimiento.

Algo similar podría ocurrir con sistemas automatizados destinados al cálculo de plazos procesales. Una configuración defectuosa que omita días inhábiles o interprete incorrectamente calendarios judiciales podría provocar la presentación extemporánea de recursos o promociones procesales, y afectar directamente el acceso efectivo a mecanismos de defensa y tutela judicial.

Desde la perspectiva de los derechos humanos, y conforme a las bases convencionales para una teoría del error (expuestas líneas arriba), el punto 4 – denominado privación errónea de derechos – es el eje. Esto no significa que todos los puntos de las bases no guarden relación con el error técnico, sino que en este caso el núcleo está en dicho punto. Con todo, el error técnico también se previene bajo el amparo del artículo 14 del PIDCP.

El debido proceso exige que las decisiones que afectan derechos fundamentales se construyan sobre bases racionales. En consecuencia, si una persona resulta perjudicada por errores automatizados presentes en pruebas, o actuaciones procesales, también se debilita su posibilidad real de defensa y de impugnación efectiva.

El problema se vuelve aún más delicado cuando las autoridades depositan confianza excesiva en sistemas automatizados. La aparente precisión matemática de ciertas herramientas puede generar una presunción informal de corrección que dificulte cuestionar sus resultados, aun cuando estos contengan errores relevantes.

Precisamente por ello, el uso de inteligencia artificial en funciones jurisdiccionales exige supervisión humana constante, y mecanismos efectivos de impugnación y corrección.

3.4. El sesgo algorítmico y el derecho a la igualdad y no discriminación

Uno de los principales riesgos de la inteligencia artificial contemporánea es que muchos sistemas aprenden a partir de datos históricos producidos por sociedades que ya contienen desigualdades estructurales. Como consecuencia, las herramientas automatizadas pueden reproducir —e incluso intensificar— prácticas discriminatorias previamente existentes.

El problema resulta especialmente delicado porque muchos sistemas algorítmicos son presentados como herramientas neutrales y objetivas. Sin embargo, se reitera, la experiencia contemporánea demuestra que los algoritmos pueden incorporar prejuicios presentes en los datos históricos con los que fueron entrenados y proyectarlos posteriormente a gran escala mediante procesos automatizados.

Uno de los casos paradigmáticos es COMPAS, sistema utilizado en Estados Unidos para evaluar riesgos de reincidencia y apoyar decisiones previas al juicio. Diversas investigaciones sostuvieron que la herramienta clasificaba con mayor frecuencia como “alto riesgo” a personas afroamericanas que posteriormente no reincidían, mientras asignaba categorías de menor riesgo a personas blancas que sí reincidían después²⁸.

El problema no consistía únicamente en que el sistema pudiera equivocarse. Todos los modelos estadísticos presentan márgenes de error. Lo verdaderamente relevante era cómo dichos errores se distribuían. Cuando una herramienta afecta de manera desproporcionada a grupos históricamente discriminados, el problema trasciende la precisión matemática y entra directamente en el terreno de la igualdad y la no discriminación.

Precisamente por ello, el sesgo algorítmico plantea desafíos especialmente complejos para el derecho internacional de los derechos humanos. La discriminación contemporánea no siempre aparece mediante reglas explícitas o categorías abiertamente excluyentes. En ocasiones opera a través de variables aparentemente neutrales —como códigos postales, historiales administrativos o patrones estadísticos— que terminan afectando de manera desproporcionada a determinados sectores sociales.

El caso neerlandés del Toeslagenaffaire constituye otro ejemplo particularmente ilustrativo. Las autoridades utilizaron sistemas automatizados para detectar supuestos fraudes relacionados con subsidios de cuidado infantil. Aunque los criterios empleados parecían administrativos y neutrales, en la práctica impactaban de forma desproporcionada sobre personas migrantes y familias con doble nacionalidad. Miles de personas fueron señaladas

²⁸ Larson, J., Mattu, S., Kirchner, L. y Angwin, J., “How We Analyzed the COMPAS Recidivism Algorithm”, *ProPublica*, 23 de mayo de 2016.

injustamente como sospechosas, enfrentando endeudamiento, deterioro económico y pérdida de confianza en las instituciones públicas.²⁹

Desde la perspectiva jurídica, el sesgo algorítmico compromete de manera directa el derecho a la igualdad y a la no discriminación, reconocidos en los artículos 2 y 26 del PIDCP. En relación con las bases de la teoría del error, su vigilancia corresponde principalmente a la primera base de la teoría convencional del error, que atiende a la prevención de desviaciones estructurales en la aplicación de derechos.

No obstante, el problema excede la mera igualdad formal: también erosiona el principio de imparcialidad estatal, pues el uso de algoritmos sesgados por parte de autoridades públicas convierte discriminaciones privadas en discriminaciones institucionalizadas. Este riesgo se agrava porque la discriminación adquiere apariencia de legitimidad al presentarse bajo lenguaje técnico y estadístico. Por ello, desde la perspectiva del PIDCP, los Estados no solo deben abstenerse de utilizar sistemas explícitamente discriminatorios, sino también supervisar activamente que las herramientas automatizadas empleadas en funciones públicas no reproduzcan desigualdades incompatibles con la dignidad humana.

Finalmente, los estándares internacionales refuerzan esta preocupación. Por ejemplo, además de las normativas ya mencionadas también el artículo 14 del Convenio Europeo de Derechos Humanos, el artículo 21 de la Carta de Derechos Fundamentales de la Unión Europea han subrayado que la igualdad exige no solo ausencia de trato diferenciado explícito, sino también prevención de impactos desproporcionados derivados de prácticas aparentemente neutrales. El sesgo algorítmico, por tanto, constituye un desafío normativo reconocido globalmente y no puede ser reducido a un problema meramente técnico.

3.5. La alucinación e integridad de la información procesal

En contextos jurídicos, este tipo de fallo adquiere especial gravedad. El derecho depende profundamente de la autenticidad de las fuentes y la posibilidad de verificar afirmaciones mediante mecanismos institucionales relativamente estables. Cuando una herramienta tecnológica produce precedentes inexistentes, se compromete directamente la integridad del procedimiento judicial.

El caso *Mata v. Avianca, Inc.*, en Estados Unidos se convirtió en uno de los ejemplos más emblemáticos de este fenómeno.³⁰ Abogados utilizaron ChatGPT para localizar precedentes judiciales aplicables a una demanda civil en

²⁹ *Toeslagenaffaire* (Raad van State – Afdeling bestuursrechtspraak, 2023).

³⁰ *Mata v. Avianca, Inc.*, *op. cit.*

Estados Unidos. El sistema proporcionó casos completamente ficticios y razonamientos inexistentes. Lo más preocupante fue que la herramienta insistió reiteradamente en la autenticidad de dichos precedentes cuando se le pidió confirmación adicional³¹.

Además del famoso caso *Mata v. Avianca* en Estados Unidos, donde los abogados fueron sancionados por presentar precedentes inventados por ChatGPT, han surgido otros ejemplos en distintas partes del mundo que muestran cómo los tribunales reaccionan frente al uso irresponsable de la inteligencia artificial.

En el Reino Unido, en los asuntos *Ayinde v London*³² el problema fue muy similar. Los abogados presentaron escritos con citas jurídicas inexistentes. El tribunal, perteneciente a la *High Court of Justice, Queen's Bench Division*, consideró esta conducta podía como desacato y ordenó el pago de costas por tiempo perdido (wasted costs), además de remitir el asunto a la *Bar Standards Board* para sanciones disciplinarias.

En Australia también se vivieron episodios semejantes. En *Re Walker*³³ ante la *Supreme Court of Victoria*, una abogada utilizó IA para preparar sus alegatos de apertura y terminó citando autoridades inexistentes. El tribunal emitió una reprimenda formal, recordando que la verificación de fuentes es una obligación profesional.

Estos ejemplos muestran como el fenómeno de las “alucinaciones” de la IA no es exclusivo de un solo país. El episodio reveló la capacidad de producir falsedad con apariencia de autoridad técnica. A diferencia de muchos errores humanos tradicionales, la alucinación algorítmica puede presentarse mediante textos convincentes y formalmente correctos.

La relevancia jurídica del problema es considerable. Los sistemas judiciales descansan en buena medida sobre la confianza en la autenticidad de los documentos y de las fuentes utilizadas dentro del proceso. Si herramientas automatizadas comienzan a introducir información inexistente con apariencia legítima, se debilitan simultáneamente la racionalidad probatoria, la confianza procesal y la integridad de la función jurisdiccional.

Además, la alucinación no representa únicamente un riesgo para abogados litigantes. También puede afectar autoridades administrativas,

³¹ Un Tracker de Alucinaciones de IA en Tribunales puede consultarse en el siguiente lugar: (*Tracker de Alucinaciones de IA En Tribunales*, 2026)

³² *Ayinde v London Borough of Haringey; Al-Haroun v Qatar National Bank*, Case Nos: AC-2024-LON-003062 y CL-2024-000435 (High Court of Justice, King's Bench Division, 23 de mayo de 2025).

³³ *Re Walker* [2025] VSC 714 (Supreme Court of Victoria, 24 de noviembre de 2025).

ministerios públicos, órganos jurisdiccionales o cualquier operador jurídico que utilice herramientas generativas sin mecanismos adecuados de verificación. Cuando herramientas automatizadas comienzan a producir información inexistente con apariencia legítima, se afecta directamente la integridad de la actividad jurisdiccional.

Desde la perspectiva del PIDCP, esto compromete directamente el derecho a obtener una resolución fundada en elementos verificables y susceptibles de contradicción. Concretamente, en la segunda base de la teoría del error no solo hay un fundamento teórico, sino también un respaldo normativo y jurisprudencial claro: las decisiones jurisdiccionales deben construirse sobre información verificable, auténtica y susceptible de contradicción. De lo contrario, se compromete tanto el derecho individual de la parte afectada como la legitimidad institucional de la justicia democrática.

Además, la alucinación presenta un desafío particularmente delicado porque puede expandirse rápidamente dentro de sistemas jurídicos altamente digitalizados.

Finalmente, antes de concluir este apartado, considero importante dejar en claro que la autonomía conceptual de las alucinaciones no obedece únicamente a diferencias en el funcionamiento interno de los modelos generativos. Habrá, quien con justificación pueda considerar a las alucinaciones como una modalidad del error técnico. Sin embargo, considero que, desde una perspectiva jurídica, ambas figuras responden a naturalezas jurídicas sustancialmente distintas. Mientras el error técnico consiste en una representación defectuosa de información existente, la alucinación supone la generación de información inexistente presentada con apariencia de autenticidad. La diferencia no radica únicamente en la magnitud del error, sino en su naturaleza epistemológica. Cada categoría, por tanto, pone en riesgo una dimensión diferente de la legitimidad del ejercicio de la función jurisdiccional.

4. A manera de conclusión: hacia un modelo de responsabilidad jurídica frente a los fallos algorítmicos

La discusión sobre responsabilidad algorítmica trasciende el ámbito estrictamente tecnológico. En el fondo, plantea una pregunta más profunda sobre el futuro del constitucionalismo y de las garantías históricamente construidas para limitar el ejercicio del poder.

Durante siglos, el derecho desarrolló mecanismos destinados a controlar arbitrariedades provenientes de autoridades humanas identificables. Hoy, una parte creciente de las decisiones públicas comienza a apoyarse en sistemas automatizados capaces de procesar enormes volúmenes de información y producir resultados revestidos de aparente objetividad matemática. La verdadera administración de justicia significa protección de la dignidad humana

y de los derechos fundamentales frente a cualquier forma de poder, incluso cuando ese poder adopte ahora forma algorítmica.

El riesgo contemporáneo no consiste únicamente en que las máquinas se equivoquen. Los seres humanos también se equivocan y quizás más. Por eso existen partidarios de minimizar los errores de las máquinas y oponerse a la hipótesis de trabajo que ha guiado esta exposición.

Por ejemplo, Richard Susskind siempre ha mantenido un optimismo desbordante sobre el uso de las tecnologías en el área jurídica y ve más beneficios que peligros. Incluso, recientemente ha expresado que apenas es el comienzo y nos espera un mundo magníficamente virtual.³⁴ Esta visión se corresponde con la postura sostenida de Ray Kurzweil, uno de los más persistentes defensores de la inteligencia artificial, quien ha llegado a afirmar que se aproxima una era en la que la materia ordinaria se transformará en *computronium*, “es decir, en materia organizada según la densidad máxima de la computación.”³⁵

Sin embargo —y a pesar de aquellas posiciones— el verdadero peligro radica en que sus errores puedan adquirir legitimidad institucional precisamente por presentarse bajo el lenguaje técnico de la eficiencia y la neutralidad algorítmica, o bien por olvidar factores no atendidos por la IA. En este sentido, Kleinberg, con ideas más moderadas, advierte que “muchos otros factores influyen en esta decisión —la disuasión, la retribución, el remordimiento— que ni siquiera se miden. Sería absurdo concluir que podemos mejorar las decisiones sobre las sentencias simplemente porque predecimos la reincidencia”.³⁶

Por ello, el desafío jurídico actual no es determinar si la inteligencia artificial debe participar en actividades estatales —pues su incorporación parece ya inevitable—, sino en establecer cuáles son los límites normativos indispensables para impedir que la automatización debilite las garantías fundamentales construidas por el derecho internacional de los derechos humanos.

La pregunta de fondo no es tecnológica, sino constitucional: cómo asegurar que, incluso en escenarios crecientemente automatizados, las personas continúen siendo tratadas como sujetos de derechos y no únicamente como datos procesables dentro de sistemas algorítmicos.

³⁴ Susskind, R., *How to Think about AI: A Guide for the Perplexed*, Oxford: Oxford University Press, 2025, pp. 143–156.

³⁵ Kurzweil, R., *La singularidad está más cerca: Cuando nos fusionamos con la IA* (trad. Carlos García Hernández), Barcelona: Deusto, 2025, p. 19.

³⁶ Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. y Mullainathan, S., “Human Decisions and Machine Predictions”, *NBER Working Paper Series*, 2017, p. 8.

Pero, además, la relación entre inteligencia artificial y derechos humanos no se limita a identificar posibles afectaciones individuales. También obliga de forma inmediata a replantear las obligaciones estatales derivadas del PIDCP frente al uso de tecnologías automatizadas.

Los Estados continúan siendo responsables por las decisiones adoptadas dentro de sus estructuras institucionales, y desde luego también cuando dichas decisiones incorporen herramientas algorítmicas. La automatización no elimina el deber de respeto y garantía de los derechos reconocidos internacionalmente.

En consecuencia, la implementación de inteligencia artificial en funciones públicas exige estándares reforzados (técnicos y jurídicos) de supervisión, impugnación, transparencia y control.

Esto implica, al menos, tres deberes fundamentales.

En primer lugar, los Estados deben evaluar previamente los riesgos que determinados sistemas pueden generar sobre derechos humanos, especialmente en ámbitos sensibles como lo es el de la justicia. A esto llamaría yo la construcción de un modelo de responsabilidad frente a fallos algorítmicos. Este modelo exige distinguir nuevamente entre las diferentes categorías de error analizadas previamente. El error técnico, el sesgo algorítmico y la alucinación generan problemas distintos de supervisión, previsibilidad y atribución jurídica. En consecuencia, aunque todos pueden comprometer derechos humanos, no necesariamente producen idénticos deberes de prevención o control.

En segundo lugar, la intervención humana no puede reducirse a una validación automática o meramente formal del resultado producido por el sistema.

Finalmente, deben existir mecanismos efectivos de impugnación, corrección y reparación frente a decisiones defectuosas derivadas del uso de inteligencia artificial. La opacidad tecnológica no puede convertirse en una barrera que impida cuestionar actuaciones estatales incompatibles con derechos humanos.

Precisamente ahí radica uno de los principales desafíos contemporáneos: asegurar que la innovación tecnológica permanezca subordinada a la dignidad humana y no a la inversa.

5. Bibliografía

Alansari, A., y Luqman, H., "Large language models hallucination: A comprehensive survey", *Computer Science Review*, vol. 61, 2026, pp. 1-79.

Athalye, A., Engstrom, L., Ilyas, A. y Kwok, K., "Synthesizing Robust Adversarial Examples", en J. Dy y A. Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 284-293.

Barocas, S., y Selbst, A. D., "Big Data's Disparate Impact" (SSRN Scholarly Paper No. 2477899), Rochester, NY: Social Science Research Network, 2016. Disponible en: <https://doi.org/10.2139/ssrn.2477899>

Belloso Martín, N., "La problemática de los sesgos algorítmicos (con especial referencia a los de género). ¿Hacia un derecho a la protección contra los sesgos?", en F. H. Llano Alonso (dir.), J. Garrido Martín y R. Valdivia Giménez (coords.), *Inteligencia Artificial y Filosofía del Derecho*, Murcia: Editorial Laborum S.L., 2022, pp. 45-75. Disponible en: <https://dialnet.unirioja.es/servlet/articulo?codigo=8556280>

Belloso Martín, N., "Sobre fairness y machine learning: el algoritmo ¿puede (y debe) ser justo?", *Anales de la Cátedra Francisco Suárez*, núm. 57, Granada: Universidad de Granada, 2023, pp. 7-38. DOI: 10.30827/acfs.v57i.25250 Disponible en: <https://revistaseug.ugr.es/index.php/acfs/article/view/25250>

Bishop, C. M., *Pattern Recognition and Machine Learning*, 1.^a ed., Nueva York: Springer, 2006.

Carbonell Bellolio, F., y Fernández Blanco, C. (eds.), *Errores en la creación y aplicación del Derecho*, 1.^a ed., Madrid: Marcial Pons, 2025.

Carbonell Bellolio, F., Meroi, A., y Uscanga, A. (eds.), *El error judicial. Problemas y regulaciones*, 1.^a ed., Valencia: Tirant lo Blanch, 2023.

National Transportation Safety Board, *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018* (Highway Accident Report NTSB/HAR-19/03, PB2019-101402), Washington, D.C.: National Transportation Safety Board, 2019. Disponible en: <https://www.nts.gov/investigations/AccidentReports/Reports/HAR1903.pdf>

Ferrajoli, L. *Derecho y Razón. Teoría del garantismo penal* (trad. Perfecto Andrés Ibáñez, Alfonso Ruiz Miguel, Juan Carlos Bayón Mohino, Juan Terradillos Basoco y Rocío Cantanero Bandrés), 5.^a ed., Madrid: Trotta, 2001.

Floridi, L., "AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis",

Philosophy & Technology, vol. 38, n° 1, 2025, p. 30. DOI: 10.1007/s13347-025-00858-9 Disponible en: <https://doi.org/10.1007/s13347-025-00858-9>

Géron, A., *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2.^a ed., Sebastopol, CA: O'Reilly Media, 2019.

"Google pide perdón por confundir a una pareja negra con gorilas", *BBC News Mundo*, 2 de julio de 2015. Disponible en: https://www.bbc.com/mundo/noticias/2015/07/150702_tecnologia_google_perdon_confundir_afroamericanos_gorilas_lv

Habermas, J. *Facticidad y validez* (trad. Manuel Jiménez Redondo), 4.^a ed., Madrid: Trotta, 2005.

Kahneman, D., y Tversky, A., "Prospect Theory: An Analysis of Decision under Risk", *Econometrica*, vol. 47, n° 2, 1979, pp. 263–291. DOI: 10.2307/1914185 Disponible en: <https://doi.org/10.2307/1914185>

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., y Mullainathan, S., "Human Decisions and Machine Predictions", *NBER Working Paper Series*, 2017, pp. 1–76.

Kurzweil, R., *La singularidad está más cerca: Cuando nos fusionamos con la IA* (trad. Carlos García Hernández), Barcelona: Deusto, 2025.

Larson, J., Mattu, S., Kirchner, L., y Angwin, J., "How We Analyzed the COMPAS Recidivism Algorithm", *ProPublica*, 23 de mayo de 2016. Disponible en: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

Llano Alonso, F., *Homo Ex Machina. Ética de la inteligencia artificial y Derecho digital ante el horizonte de la singularidad tecnológica*, Valencia: Tirant lo Blanch, 2024.

Moreno Cruz, R., "El modelo garantista de Luigi Ferrajoli: Lineamientos generales", *Boletín Mexicano de Derecho Comparado*, vol. 40, n° 120, 2007, pp. 825–852.

Moreno Cruz, R., "Jürgen Habermas. Baedeker de su propuesta jurídica", *Problema. Anuario de Filosofía y Teoría del Derecho*, n° 17, 2023, pp. 221–250. DOI: 10.22201/ijj.24487937e.2023.17.18215 Disponible en: <https://doi.org/10.22201/ijj.24487937e.2023.17.18215>

Saarenpää, A., "Law as Communication in a Digital Constitutional State", en E. Schweighofer, H. Zech, F. Kummer y A. Saarenpää (eds.), *Human-Machine Cooperation in Cyberspace*, Bern: Weblaw, 2026, pp. 29–39.

Solar Cayón, J.I., *Inteligencia artificial jurídica e imperio de la Ley*, Valencia: Tirant lo Blanch, 2025.

Susskind, R., *How to Think about AI: A Guide for the Perplexed*, Oxford: Oxford University Press, 2025.

Tracker de Alucinaciones de IA en Tribunales, "Tracker de Alucinaciones de IA en Tribunales", 2026. Disponible en: <https://tracker-de-alucinaciones.lovable.app/>

Tuzet, G., *Filosofía de la prueba jurídica*, 1.^a ed., Madrid: Marcial Pons, 2021.

Casos

Ação Trabalhista Ordinária (ATOrd) nº 0001062-55.2025.5.08.0130 (3ª Vara do Trabalho de Parauapebas, 6 de mayo de 2026). Disponible en: <https://www.diariojudicial.com/uploads/0000061635-original.pdf>

Ayinde v London Borough of Haringey; Al-Haroun v Qatar National Bank Case Nos: AC-2024-LON-003062 y CL-2024-000435 (High Court of Justice, King's Bench Division, 23 de mayo de 2025).

Facultad de Atracción 2/2012 (Primera Sala de la Suprema Corte de Justicia de la Nación, 2012). Disponible en: https://www2.scjn.gob.mx/juridica/engroses/1/2012/69/2_137627_1275_firmado.pdf

Mata v. Avianca, Inc. 1:2022cv01461 (United States District Court for the Southern District of New York, 22 de junio de 2023). Disponible en: <https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/>

Mathews v. Eldridge 424 U.S. 319 (U.S. Supreme Court, 1976). Disponible en: <https://supreme.justia.com/cases/federal/us/424/319/>

Re Walker [2025] VSC 714 (Supreme Court of Victoria, 24 de noviembre de 2025). Disponible en: https://websitedc.s3.amazonaws.com/documents/Re_Walker_Australia_24_November_2025.pdf

Toeslagenaffaire (Raad van State - Afdeling bestuursrechtspraak, 2023). Disponible en: <https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RVS:2023:3620>

Principios y observaciones

ONU *Conjunto de principios para la protección y la promoción de los derechos humanos mediante la lucha contra la impunidad*, 2005. Disponible en: <https://docs.un.org/es/E/CN.4/2005/102/Add.1>

ONU Observación general N° 32: Artículo 14. El derecho a un juicio imparcial y a la igualdad ante los tribunales y cortes de justicia, 2007. Disponible en: <https://www.refworld.org/es/leg/coment/ccpr/2007/52583>

UNESCO *Directrices para el uso de sistemas de IA en Cortes y Tribunales*, 1.^a ed., París: UNESCO, 2026.

Pactos y Leyes

Pacto Internacional de Derechos Civiles y Políticos, 1966.

Unión Europea, *Ley de Inteligencia Artificial (AI Act)*, 2024. Disponible en: <https://ai-eu-act.eu/es/>