Inteligencia artificial explicable (AIX) y prevención de daños

por Matilde Pérez(*)

Sumario: I. Introducción. – II. Las cajas negras o *black box*. – III. Salir de la opacidad. La búsqueda de una IA explicable. – IV. ¿Qué es la Inteligencia Artificial Explicable (XAI)? – V. Principios aplicables a la IAE. – VI. La IAE como instrumento de gestión de riesgos ante la incertidumbre. – VII. Herramientas y métricas para una inteligencia artificial más confiable. – VIII. Conclusión.

I. Introducción

En ocasión de celebrar el 10º aniversario de la entrada en vigor del Código Civil y Comercial, la organización de las XXX Jornadas Nacionales de Derecho Civil continúa la línea de las cuatro últimas jornadas respecto del derecho de daños y las tecnologías.

Es en este sentido que la vorágine de las innovaciones tecnológicas requiere de los operadores jurídicos especial atención en lo que hace a áreas tan sensibles como la libertad, la autonomía de la voluntad o la centralidad de la dignidad humana.

En esta Era 4.0 la colaboración y transversalidad entre el Derecho y otras ciencias es una realidad que trasunta todas las áreas del conocimiento y hace que las tecnologías basadas en sistemas de IA deban ser analizadas desde el desarrollo de su diseño hasta su posterior utilización.

En esta oportunidad, el abordaje propuesto por las XXX Jornadas de Derecho Civil en la Comisión 3 es el de "Daños derivados de productos, servicios y bienes digitales" afirmación que coloca a estas nuevas realidades en la órbita jurídica de la responsabilidad civil.

En la vida cotidiana, los productos digitales son bienes intangibles que se producen, almacenan y distribuyen en formato digital. El software, los libros electrónicos, música, videos, cursos, plantillas, entre muchos otros, se caracterizan por su inmediatez, su fácil uso y formatos accesibles.

Los servicios digitales son actividades o funcionalidades ofrecidas a través de plataformas o redes digitales. Almacenamiento en la nube, redes sociales, software como servicios (SaaS) o programas de streaming requieren de una conexión de acceso a internet, pueden ser personalizadas de acuerdo con las necesidades del usuario y, por lo general, se obtienen por suscripciones sean pagas o gratuitas.

El concepto de bienes digitales comprende aquellos objetos virtuales que se pueden comprar, vender o coleccionar en línea. Se representan, por lo general, en tokens no fungibles (NFT) y en criptoactivos. Entre los más conocidos, las criptomonedas, activos digitales, arte digital o el mundo del metaverso.

Se percibe de manera más o menos consciente la existencia de estas realidades, así como la presencia de la inteligencia artificial en los distintos sistemas que se valen de ella o servicios y productos que se desarrollan o utilizan la IA, que transforman la sociedad, las formas de pensar, de relacionarse o cómo realizar las tareas más simples o complejas.

Si esa percepción se transforma en pregunta acerca de cómo la IA funciona, cómo se crea, cómo se aplica, quiénes están detrás o qué pasa si la IA se equivoca o no sabe, la mayoría ignora o tiene un conocimiento somero o superficial al respecto.

Quizás, en muchos casos, la persona se queda con la cara amable en el ahorro de tiempo, la disponibilidad de información o el acceso a bienes y servicios. Sin embargo, hay situaciones en áreas como la sanidad o el uso de datos en que parecen mostrar lo contrario, pues se presentan situaciones como discriminación, sesgos negativos o tomas de decisiones por estos sistemas difíciles de explicar o comprender.

Y es que estos modelos de IA impactan en un amplio abanico de cuestiones tan dispares como la solicitud de un préstamo, el dictado de un acto administrativo, una praxis médica o el ingreso universitario.

La propuesta de este trabajo es analizar posibles soluciones a los problemas conocidos como de *cajas negras* o *black box* en los que hay una dificultad de interpretar y comprender cómo un modelo de IA llega a una determinada conclusión que se verá reflejada en el espectro de los productos, servicios o bienes digitales, con la posibilidad de generar daños de entidad y extensión diversas.

Esa búsqueda se enmarca en la función preventiva del daño a través de sus principios integradores, a los que se agrega la posibilidad de pensar en una IA explicable (IAE o AIX) o cómo desde el diseño del producto se puede anticipar la existencia de dañadores tecnológicos y así contribuir a la seguridad jurídica, la protección de los usuarios, los consumidores y los proveedores, así como ser parte de la gobernanza de los Estados.

II. Las cajas negras o *black box*

Existen sistemas algorítmicos basados en lenguajes de aprendizaje automático o *machine learning* en los que no se sabe con certeza cuáles son los mecanismos que realizan para llegar a un resultado, esto es, las llamadas cajas negras o *black box*.

Estos fenómenos se engloban dentro de los defectos de diseño y, por tanto, con potencialidad nociva desde su desarrollo inicial.

En estas situaciones, el sistema actúa como una red neuronal en la que es muy difícil poder determinar cómo se llega a determinados resultados.

Este aprendizaje automático presenta varios modelos: a) *Supervised learning* o aprendizaje supervisado en el que el sistema está entrenado con conjuntos de datos estructurados para predecir resultados de manera precisa. b) *Reinforcement learning* o aprendizaje de refuerzo que es un sistema entrenado a base de prueba de error sin contar con un conjunto de datos. c) *Deep learning* o aprendizaje profundo. El sistema debe ser entrenado en un conjunto masivo de datos, millones de datos; este entrenamiento puede ser estructurado o no estructurado, sin necesidad de intervención humana para el procesamiento de tales datos. d) Redes neuronales profundas. El diseño de diagramas neuronales cada vez más extensos requiere de un sistema que permita garantizar la transparencia de tales procesos.

En el *machine learning* se da la posibilidad de "aprender" sin haber programado de una manera explícita. Se basa en la identificación de patrones sobre los *inputs* y la aplicación subsiguiente del "conocimiento".

A mayor autonomía del sistema se advierte una mayor falta de transparencia y accesibilidad con relación a las predicciones y soluciones a las que llegan estos sistemas.

El estado de los conocimientos científicos se topa con nuevos contextos de incertidumbre que nacen de cómo se autoprocesan los algoritmos y se abocan a sus soluciones;

Nota de Redacción: Sobre el tema ver, además, los siguientes trabajos publicados en El Derecho: *Las "tecnologías reproductivas" y la ética médica*, por Elisabet Agustina Vidal, ED, 259-913; *Responsabilidad civil en internet: avance de las nuevas tecnologías de la información y asignaturas pendientes del sistema jurídico*, por Marcelo Oscar Vuotto, ED, 261-860; *El nuevo Código Civil y Comercial y el rol de nuestra formación jurídica*, por Mario A. Zinny, ED, 263-870; *El Código Civil y Comercial en clave de derechos humanos. El impacto del derecho internacional de los derechos humanos en la aplicación e interpretación del nuevo derecho privado argentino*, por Marcelo Trucco, ED, 264-810; *El uso de la tecnología y la gestión de la comunicación en la mediación actual*, por Juan Fernando Gouvert, ED, 275-771; *El derecho ante la inteligencia artificial y la robótica*, por Verónica Elvia Melo, ED, 276-493; *La protección de los datos personales en internet (una tarea ineludible)*, por Esteban Ruiz Martínez, ED, 284-726; *La comunidad humana en la era tecnológica*, por Leonardo Pucheta, ED, 282-1044; *Robótica e inteligencia artificial: nuevos horizontes de reflexión*, por Leonardo Pucheta, ED, 283-925; *Los paradigmas del derecho privado codificado. El caso argentino: de persona a individuo*, por Gabriel F. Limodio, ED, 286-461; *El concepto de persona frente a las tecnologías disruptivas: persona humana, persona jurídica, ¿persona electrónica?*, por Verónica Elvia Melo, ED, 289-1386; *Derecho de los robots*, por Pilar Moreyra, ED, 291-708. Todos los artículos citados pueden consultarse en www.elderechodigital.com.ar.

(*) Doctora en Ciencias Jurídicas. Especialista en Derecho Administrativo. Directora del Centro de Innovación Jurídica UCA. Profesora titular de las asignaturas Obligaciones Civiles y Comerciales, Derecho de Daños y Derechos Reales en la Facultad de Derecho de la Universidad Católica Argentina. Profesora en el Doctorado en Ciencias Jurídicas y en la Maestría de Derecho Civil Patrimonial, miembro del Comité Asesor del Doctorado en Ciencias Jurídicas y miembro de la Comisión de Abogacía Digital (UCA). Profesora invitada en Universidades nacionales y extranjeras. Subdirectora del Suplemento "Derecho, Innovación y Desarrollo Sustentable" en Editorial El Derecho. Autora de libros, capítulos de libros y ponencias. Correo electrónico matildeperez@uca.edu.ar. Código ORCID 009-0008-2189-701X.

sin embargo, la probabilidad dañosa existe, está presente el "humo de peligro" que debe mover a todos los operadores a canalizar alternativas para lograr que esos efectos nocivos derivados de las decisiones adoptadas en estos contextos impacten lo menos posible.

En idéntico sentido, al determinarse la responsabilidad por los daños causados por el uso de estas tecnologías, no puede argumentarse como eximente de la responsabilidad la existencia de un caso fortuito interno o la falta de conocimiento científico.

III. Salir de la opacidad. La búsqueda de una IA explicable

Esas mejoras reconocen en la trazabilidad y la transparencia dos mecanismos esenciales tanto desde la perspectiva del estudio de la relación costo-beneficio como desde la función preventiva de daños, los servicios de atención al cliente y la disminución de efectos negativos de carácter masivo como aquellos que provienen de los casos fortuitos internos, respecto a la transparencia, seguridad y confiabilidad de las decisiones automatizadas.

Ante este panorama, la *inteligencia artificial explicable* adquiere una relevancia primordial, puesto que permite acercarse al procesamiento de datos interno de los algoritmos para tornarlos comprensibles y evitar decisiones discriminatorias o perjudiciales, por lo que es una herramienta que encuadra en las disposiciones de los arts. 1710, 1711 y 1712 del CCC.

Los algoritmos basados en IA, en especial, las redes neuronales profundas tienen una mayor presencia en cada tarea cotidiana debido al incremento del uso del aprendizaje profundo o *deep learning*.

En sectores como la salud, finanzas y justicia, la IA explicable garantiza que las decisiones automatizadas sean comprensibles y justificadas, evitando daños derivados de interpretaciones erróneas o falta de transparencia[1].

IV. ¿Qué es la Inteligencia Artificial Explicable (XAI)?

La inteligencia artificial explicable (IAE o XAI por su sigla en inglés) se orienta a transparentar los procesos internos y las lógicas de decisión de los algoritmos. Esto implica que, en lugar de operar como *cajas negras*, los sistemas de IA deben ofrecer mecanismos de interpretación que posibiliten a desarrolladores, usuarios y autoridades comprender, auditar y supervisar sus decisiones.

La explicabilidad se torna en un principio que busca entender y comprender cómo los sistemas llegan a esas decisiones.

El término IAE fue acuñado por el Proyecto DARPA[2] en 2016 destinado a solucionar la falta de transparencia y el entendimiento de los sistemas de IA, en especial, en aplicaciones militares.

Promueve un conjunto de herramientas tecnológicas y algorítmicas que puedan brindar explicaciones de alta calidad, interpretables e intuitivas para los seres humanos.

Esa explicabilidad debe estar dirigida a asegurar que la decisión adoptada por estos sistemas no es sesgada, es antidiscriminatoria y los procesos de carga de datos están atravesados por la buena fe.

Se pretende evitar la existencia de una relación de causalidad entre la *caja negra* y el daño, dado que al ser explicable las personas pueden evaluar cómo sus datos son utilizados y tomar decisiones acerca de ellos o de su privacidad.

Aun así, pueden existir brechas o grietas acerca de la manera en que la transparencia y la explicabilidad afectarán la confianza pública o si tales sistemas pueden ser dirigidos de manera dolosa como una mera formalidad.

Es por ello que la IAE debe partir de la detección de sesgos algorítmicos que son considerados como errores sistemáticos y estructurados en un sistema de IA que generan desigualdades y resultados injustos. La existencia de *cajas negras* es, como se dijo, un defecto de diseño que ni los desarrolladores pueden explicar con fiabilidad porque se desconoce cómo se llega a ese sesgo.

La IAE tiene que abrir un camino para facilitar el reconocimiento y la concientización precisa de posibles sesgos entre las diversas partes interesadas en ese ecosistema de IA para lograr la anticipación o mitigación de estos agentes con potencialidad dañosa.

Ir hacia *cajas blancas* o *cajas de cristal* es parte del derecho a la explicación por el que la persona titular de datos tiene el derecho a la explicación de la decisión tomada después de la evaluación algorítmica, esto es, la información significativa sobre la lógica involucrada en las decisiones automatizadas[3].

En el caso de las empresas ubicadas en la jurisdicción de la Unión Europea, solo pueden tomar determinadas decisiones siempre que puedan brindar explicaciones de cómo se llega a esa decisión para que sean más visibles y fáciles de detectar al revelar qué está haciendo el sistema y cómo llega a sus decisiones[4].

En esta búsqueda de *cajas blancas*, la explicabilidad es una necesidad inherente a los seres humanos que se vincula con la percepción y la cognición, esto es, una expresión de que todos los seres humanos por naturaleza tienden al saber.

Y es allí donde acude la puesta en marcha de estudios de espectro de usuarios donde la experiencia personal sustituya o complemente a las técnicas de computación o evaluaciones basadas en la funcionalidad, en especial, en materia de sesgos y discriminación. Las investigaciones sin referencias en lo humano no logran contextualizar la IAE como mecanismos para eliminar la opacidad de los modelos.

Esa incorporación de la perspectiva humana se centra, primero, en la incorporación de expertos como desarrolladores de IA o usuarios profesionales como en el caso de los profesionales de la salud con relación a las herramientas de IA para el diagnóstico, en áreas de cibercrimen o en el desarrollo de sistemas para la toma de decisiones judiciales, a modo de ejemplo[5].

No obstante, pareciera que los usuarios que no son expertos son los que se pueden ver más afectados por las decisiones de la IA y con preocupaciones legítimas sobre trato injusto por parte del sistema.

Este es un punto clave para que la transparencia, la responsabilidad y la explicabilidad de la IA se centre en el ser humano como viene propugnándose en los diversos textos internacionales.

Los estudios de usuarios en materia de IAE deben partir de la *interpretabilidad*, esto es, la capacidad del sistema para explicar o presentar predicciones del modelo en términos que puedan ser entendidos por las personas.

Las explicaciones comprensibles, las interfaces accesibles y las interacciones centradas en el usuario permiten convertir a la IAE en una herramienta de ayuda a los usuarios para comprender por qué se deriva –o no– un resultado de un sistema, cuándo puede tener éxito o fallar o cuándo confiar en el sistema o saber que cometió un error[6].

V. Principios aplicables a la IAE

El Instituto Nacional de Estándares y Tecnología (NIST) de Estados Unidos establece cuatro principios claves de la IAE como método para garantizar que los sistemas de IA sean transparentes, comprensibles y confiables para los usuarios[7].

(1) González-Arencibia, M.; Ordoñez-Erazo, H.; González-Sanabria, J., "Explainable Artificial Intelligence as an Ethical Principle", Rev. Ingeniería, N° 29, vol. 2, dic. 2024. Disponible en https://doi.org/10.14483/23448393.21583, consultado el 05/06/2025.

(2) DARPA. Defense Advance Research Project Agency. Información adicional disponible en https://onlinelibrary.wiley.com/doi/epdf/10.1609/aimag.v40i2.2850, consultado 04/06/2025.

(3) Selbst, A.; Powles, J., "Meaningful information and the right to explanation", *Proceedings of 1st. Conference on Fairness, Accountability and Transparency*, PMLR, 81, p. 48, 2018. Disponible en selbst18a.pdf, consultado el 05/06/2025.

(4) Y. Rong *et al.*, "Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2104-2122, abril 2024. Disponible en https://ieeexplore.ieee.org/abstract/document/10316181, consultado el 05/06/2025.

(5) Ching-Hua, Ch., Ruoyu, S.; Shiyun, T.; Wan-Hsiu, S., "EXplainable Artificial Intelligence (XAI) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives", Telematics and Informatics, Revue, Volume 91, 2024, 102135. Disponible en https://www.sciencedirect.com/science/article/pii/S073658532400039X, consultado el 05/06/2025.

(6) Gunning, D.; Aha, D., "DARPA's Explainable Artificial Intelligence (XAI) Program", AI Magazine, 40(2), 2019, pp. 44-46. Disponible en https://doi.org/10.1609/aimag.v40i2.2850, consultado el 05/06/2025.

(7) Phillips, J.; Hahn, C. *et al.*, *NISTIR 8312 Four Principles of Explainable Artificial Intelligence*. National Institute of Standards and Technology, septiembre 2021. Disponible en https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf, consultado el 05/06/2025.

1. Transparencia y comprensibilidad: A través de ellos se puede ir hasta el diseño del sistema con el objeto de poder detectar errores, sesgos o comportamientos no previstos.

2. Detección anticipada de riesgos: En los procesos de toma de decisiones automatizadas la gestión de los riesgos permite abaratar costos directos, pero también evitar daños presentes como la detección de sesgos y daños futuros generados por el uso indebido de datos personales por estos sistemas, riesgos de desarrollo de producto detectado como daño tardío o, a través de la información y el lenguaje claro, contribuir a que los ciudadanos conozcan cómo operar determinados sistemas.

3. Responsabilidad y consentimiento informado: Con la transparencia proporcionada por la XAI, tanto reguladores como responsables internos pueden evaluar críticamente los procesos automatizados, lo que favorece la rendición de cuentas y la adaptación de medidas correctivas en tiempo real.

4. Fortalecimiento de la confianza: Al hacer comprensible el algoritmo, se fomenta la confianza de todos los actores involucrados –desde clientes hasta organismos reguladores– en que los sistemas se gestionan de manera ética y segura.

VI. La IAE como instrumento de gestión de riesgos ante la incertidumbre

En el derecho de daños, el análisis de los riesgos y su gestión en contextos de incertidumbre permite acceder al control de la causalidad y, así, controlar la posibilidad de producción de daños masivos.

El despliegue de sistemas de IAE contribuye a identificar anomalías o sesgos discriminatorios en fases tempranas de los ciclos de IA, lo que permite la puesta en marcha de estrategias que tiendan a evitar la opacidad de los modelos.

Posibilita la evaluación continua y la adaptación a las nuevas normativas por cuanto la transparencia que hace a la esencia de la IAE facilita una revisión constante automatizada y humana de los algoritmos, lo que permite actualizar protocolos de actuación, marcos de autorregulación y adaptaciones legislativas.

El hacer públicos los mecanismos de decisión es un aliciente para los desarrolladores y operadores para procurar una mejora en los estándares de integridad y seguridad acudiendo a medidas de prevención y técnicas de *compliance*, lo que también puede contribuir a resultados más satisfactorios en auditorías y evaluaciones externas.

En paralelo, se presentan varias dificultades.

La complejidad técnica en el desarrollo de algoritmos potentes y transparentes requiere de mayores inversiones en investigación, desarrollo y capacitación que no siempre están en condiciones de ser soportadas o sus costos se trasladan a los productos y servicios.

La vertiginosidad de los sistemas de IA débil y fuerte, la competencia del mercado, la falta de estándares universales y marcos normativos tan dispares como el Reglamento General de Protección de Datos o el Reglamento de IA de la Unión Europea o la aplicación de normativas generales atentan contra una IAE segura y responsable.

Desde los aspectos computacionales pueden presentarse situaciones en las que la explicabilidad puede comprometer la eficiencia o la precisión del sistema, por lo cual debe establecerse el balance entre un diseño explicable de la IA y la puesta en marcha del sistema en forma autónoma o como soporte.

Es entonces que la colaboración entre los organismos públicos, el sector privado y los expertos se erige como uno de los baluartes en la búsqueda de mecanismos sin opacidades en sistemas complejos.

VII. Herramientas y métricas para una inteligencia artificial más confiable

Desde hace más de cien años es una constante el recurrir a diversos mecanismos de *soft law* y, dentro de ellos, la autorregulación que desarrolla y armoniza un sistema de referencias propias y vinculantes para quienes las suscriben frente a la burocracia o la inoperancia estatal[8].

Bajo este paraguas de la autorregulación, el Observatorio de Políticas de Inteligencia Artificial de la OCDE recopila y comparte diversas herramientas y métricas diseñadas para ayudar a los diversos actores de la IA a desarrollar sistemas confiables que respeten los derechos humanos, que sean transparentes, explicables y seguros[9].

Se destacan dos metodologías:

1. Local Interpretable Model-agnostic Explanations (LIME) permite aproximar modelos complejos con versiones más simples y que se puedan interpretar para explicar predicciones de tipo individual de una manera rápida.

2. SHapley Additive exPlanations (SHAP) que cuantifica la contribución exacta de cada variable de entrada a una predicción para lo que utiliza principios de la gamificación o teoría de los juegos. Es mucho más precisa que la anterior, pero tiene un costo mayor.

Uno de los ámbitos en los que se procura la implementación de estos modelos para la IAE es el de diagnóstico médico que puede contribuir, por ejemplo, a interpretar que la edad y la presión arterial de un paciente pueden ser los principales factores que concluyen en un diagnóstico de algo riesgo de infarto, aunque SHAP explicaría de manera exacta cómo contribuye cada variable.

En el caso de una entidad bancaria, se puede implementar un sistema de aprendizaje profundo para la detección de fraude, y usar valores SHAP para auditar sus decisiones y garantizar que no se toman decisiones discriminatorias.

Existen otras propuestas que combinan el uso de datos abiertos con sistemas de IAE debido a su interconectividad, lo que permite entonces potenciar la transparencia, la confianza y la responsabilidad en la toma de decisiones basadas en IA.

VIII. Conclusión

La *inteligencia artificial explicable* representa una herramienta poderosa para la gestión de riesgos, en tanto que su capacidad para clarificar y verificar el proceso de toma de decisiones se alinea estrechamente con el *principio de precaución*. Este enfoque permite anticipar vulnerabilidades y prevenir daños en entornos donde la incertidumbre y la complejidad tecnológica pueden desembocar en consecuencias indeseadas.

De cara al futuro, es crucial fomentar la colaboración interdisciplinaria y el desarrollo normativo que encauce la innovación hacia prácticas seguras, éticas y transparentes. La sinergia entre XAI y el principio de precaución no solo fortalece la confianza del usuario, sino que también contribuye a construir sistemas de IA que, lejos de ser cajas negras, sean mecanismos abiertos al escrutinio y la mejora continua.

VOCES: PERSONA - TECNOLOGÍA - INFORMÁTICA - TRATADOS INTERNACIONALES - DERECHOS HUMANOS - CÓDIGO CIVIL Y COMERCIAL - DERECHO CIVIL - RESPONSABILIDAD CIVIL - DAÑOS Y PERJUICIOS - INTELIGENCIA ARTIFICIAL - ORDEN PÚBLICO - PERSONAS JURÍDICAS - PRINCIPIOS GENERALES DEL DERECHO - INTERNET - PODER JUDICIAL - DERECHOS Y GARANTÍAS CONSTITUCIONALES - RESPONSABILIDAD DEL GUARDIÁN - RESPONSABILIDAD OBJETIVA

(8) Pérez Álvarez, M., *El principio de precaución y los riesgos de desarrollo*, Buenos Aires, El Derecho, 2024, pp. 143-155.

(9) OECD.AI. *Policies, data and analysis for trustworthy artificial intelligence*, julio 2024. Disponible en https://oecd.ai/en/, consultado 05/06/2025.