REVISTA CHILENA
DE DERECHO Y TECNOLOGÍA

VOL. 14 (2025) • e74040 • DOI 10.5354/0719-2584.2025.74040
RECIBIDO: 3/3/2024 • APROBADO: 2/12/2025 • PUBLICADO: 31/12/2025

DOCTRINA

La inteligencia artificial explicable como fundamento para su gobernanza: Implicancias en la responsabilidad civil

Explainable artificial intelligence as a basis for its governance: Implications for civil liability

Darío Parra Sepúlveda
Universidad Austral de Chile

Ricardo Concha Machuca
Universidad de Concepción, Chile

RESUMEN En este artículo se analiza la explicabilidad, en primer lugar, como requisito de la fiabilidad, principio ético que sirve de fundamento para la gobernabilidad de la inteligencia artificial. Desde esta perspectiva, se analizan los conceptos de gobernabilidad, fiabilidad y explicabilidad de la inteligencia artificial (IA), al hilo de su evolución, principalmente en el marco de la Unión Europea. En segundo lugar, se estudia si la inteligencia artificial explicable es relevante o no para los regímenes de responsabilidad por los daños y perjuicios ocasionados por los sistemas de IA. Ello nos permite reflexionar sobre la importancia de incorporar la explicabilidad en los nuevos esquemas de responsabilidad que se generen para abordar los problemas generados en materia de daños por la IA.

PALABRAS CLAVE Inteligencia artificial explicable, gobernanza de la IA, responsabilidad civil, inteligencia artificial avanzada, desarrollo tecnológico.

ABSTRACT This article analyzes explainability, firstly, as a requirement for reliability, an ethical principle that serves as a foundation for the governance of artificial intelligence. From this perspective, the concepts of governance, reliability and explainability of artificial intelligence (AI) are analyzed, in line with their evolution, mainly in the framework of the European Union. Secondly, we study whether or not explainable artificial intelligence is relevant to liability regimes for damages caused by AI systems. This allows us to reflect on the importance of incorporating explainability in the new liability

schemes that will be generated to address the problems generated in terms of damages caused by AI.

KEYWORDS Explainable artificial intelligence, AI governance, civil liability, advanced artificial intelligence, technological development.

Introducción

Cada vez tenemos un mayor grado de interacción con la inteligencia artificial (IA). El progresivo desarrollo de sistemas algorítmicos para la realización de tareas, que hasta hace poco tiempo eran realizadas solo por seres humanos, ha provocado un amplio debate ético y jurídico sobre las implicancias de la IA en nuestra vida. Ello ha dado lugar a un proceso creciente de adaptación del ordenamiento jurídico a los diversos problemas que estamos teniendo en nuestra convivencia cada vez más intensa con las tecnologías de IA (Solar, 2020; Parra y Concha, 2021: 2).

De acuerdo con Narváez (2019: 211), la relación de la inteligencia artificial con el derecho se ha encaminado desde dos perspectivas: i) el desarrollo de un marco normativo y de gobernanza de la IA a nivel mundial; y ii) la aplicación propia de la IA a través de sistemas o programas para diversas áreas de la disciplina jurídica. Este estudio se enmarca en la primera de estas cuestiones. Por tanto, no analizamos aquí las tensiones que se generan con el uso de la inteligencia artificial como herramienta para resolver cuestiones propiamente jurídicas, como sucede con las bases de datos legales y de jurisprudencia, la redacción de contratos o el uso de sistemas expertos jurídicos en la resolución de casos judiciales (Martínez, 2012: 833-834). Nuestro trabajo se sitúa en la necesidad de adaptar los marcos normativos y de gobernanza para responder a relaciones cada vez más complejas que se derivan del uso intensivo de la inteligencia artificial en los diversos ámbitos de nuestra vida.

La tecnología IA posee un increíble potencial para transformar nuestra forma de vivir y trabajar en las próximas décadas (Comisión Europea, 2021a: 1). De acuerdo con Dafoe (2018: 8), la IA avanzada (sistemas avanzados de inteligencia artificial) podría jugar un papel fundamental en la resolución de los problemas existentes a nivel global —el cambio climático o los conflictos internacionales— o ayudarnos a mejorar drásticamente la salud, la sostenibilidad, la ciencia y la comprensión de nosotros mismos. No obstante, toda esta potencialidad de mejoras en nuestra calidad de vida está generando riesgos y consecuencias jurídicas no previstas en nuestros ordenamientos, obligando a una profunda revisión de nuestros esquemas normativos para dar respuesta a los múltiples desafíos que la IA plantea al derecho.

Retomando las ideas de Dafoe (2018: 8), estos riesgos constituyen externalidades negativas que pueden afectar a las personas y que son especialmente difíciles de ges-

tionar. La creación de instituciones y marcos jurídicos adecuados es una condición indispensable para abordar adecuadamente estos riesgos. En este contexto, la comunidad internacional ha comenzado a reconocer la necesidad de avanzar hacia una gobernanza de la IA que debería propiciar una transición adecuada hacia un mundo con sistemas de inteligencia artificial avanzados. De esta forma, en los últimos años, se ha consolidado el consenso en torno a un conjunto de principios éticos y directrices globales sobre IA que tienen reconocimiento jurídico, como la transparencia, la justicia y la equidad, el respeto a los derechos humanos y los valores democráticos, la responsabilidad y la privacidad (Jobin, Ienca y Vayena, 2019). No obstante, a pesar de ir asentándose acuerdos fundamentales sobre la forma de diseñar y utilizar la inteligencia artificial, subsisten los desafíos sobre cómo implementar tales principios, hacerlos respetar y repartir los costes y beneficios de la digitalización (Taeihagh, 2021).

Como se verá en los siguientes apartados, el debate jurídico muestra que las repuestas desde el derecho no pueden limitarse a ajustar categorías dogmáticas y normativas preexistentes, sino que debe incorporar los dilemas de diseño institucional y de política pública que plantea una tecnología global, de rápida expansión y difícilmente controlable. En el ámbito normativo, la Unión Europea ha dado un paso decisivo en la gobernanza de la IA con la aprobación del Reglamento (UE) 2024/1689, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Artificial Intelligence Act), que establece un marco de regulación basado en el riesgo para los sistemas de inteligencia artificial y que tiene por finalidad la garantía de una IA fiable y centrada en el ser humano (artículo 1) como estándar para la Unión Europea.

En este contexto, nuestro estudio se enmarca en el análisis de la explicabililidad de la inteligencia artificial desde dos dimensiones complementarias: i) como característica de un marco normativo general y de gobernanza que fomente una IA fiable y respetuosa de los derechos de las personas; y ii) en cuanto condición para avanzar hacia un régimen de responsabilidad civil de la IA claro y efectivo. Desde la primera perspectiva, el análisis se hará a partir de los avances que ha realizado la Unión Europea en la elaboración de un marco general normativo y de gobernanza de la inteligencia artificial, por tratarse de uno de los sistemas jurídicos que mayor esfuerzo ha dedicado a este objetivo.

La explicabilidad es uno de los requisitos para lograr que la inteligencia artificial sea fiable. Este último concepto representa la necesidad de construir un ecosistema de confianza para que las personas, las empresas y las administraciones desarrollen, utilicen y adopten soluciones basadas en inteligencia artificial (Comisión Europea, 2021a). Las directrices éticas para una IA fiable propuestas por el grupo de expertos de alto nivel sobre inteligencia artificial de la Comisión Europea establecen que la fiabilidad es un requisito previo para el uso de la inteligencia artificial, que integra tres aspectos esenciales: la IA debe ser lícita, ética y robusta (Grupo Independiente

de Expertos de Alto Nivel sobre Inteligencia Artificial o AI HLEG, 2019: 6). De esta forma, una IA fiable debe estar centrada en el ser humano, guiarse por el respeto a los derechos fundamentales y los principios democráticos, tener solidez técnica y desarrollarse exclusivamente en ámbitos lícitos (AI HLEG, 2019; Antonov, 2022).

En el mismo sentido, Aneja (2021: 29) señala que «la evidencia creciente de los múltiples riesgos y perjuicios que comportan los sistemas de toma de decisiones mediante algoritmos han urgido la cuestión de la gobernanza de la inteligencia artificial». Del mismo modo, Robles (2020: 3-4) expresa que la falta de un marco jurídico general para la IA —tanto a nivel internacional como a nivel nacional— ha obligado a recurrir a la invocación de los principios o de los estándares técnicos para dar respuestas a los diversos problemas jurídicos que se están suscitando con la IA. En línea con lo anterior, Taeihagh (2021: 138-139) advierte que la gobernanza de la IA enfrenta desafíos multifacéticos, que incluyen cuestiones como la capacidad regulatoria de los Estados, la coordinación internacional y la interacción entre normas duras, *soft law* y estándares técnicos.

Todo ello, como hemos anticipado, en los últimos años ha comenzado a trabajarse —por parte de la comunidad internacional, de algunas organizaciones internacionales, como la Unión Europea, y en menor medida en países como Estados Unidos o China— en el desarrollo de un marco de gobernanza para los sistemas de inteligencia artificial, que permita aprovechar al máximo el potencial de una IA fiable (Antonov, 2022: 42-43). Ello va dando vida a un marco normativo fragmentado, en el que coexisten iniciativas de *soft law*, estrategias nacionales, marcos regionales vinculantes o propuestas de tratados internacionales, que progresivamente van moldeando un «estatuto universal» de la inteligencia artificial que no termina de configurarse, aunque sí se van decantando principios jurídicos que promueven la fiabilidad, la transparencia, la rendición de cuentas y la explicabilidad de la IA (Corrêa y otros, 2023).

Pero la IA explicable no solo es un requisito para conseguir una IA fiable, sino que también es una condición para avanzar hacia un régimen de responsabilidad civil de la inteligencia artificial claro y efectivo. Este es el segundo ámbito de estudio de este análisis. No nos centraremos en las respuestas que los ordenamientos jurídicos han ido dando a los diversos problemas que se están presentando en relación con la responsabilidad civil ante acciones provocadas por el uso de máquinas dotadas de inteligencia artificial,[1] sino que revisaremos la relevancia que tiene la IA explicable para avanzar hacia un derecho de daños por IA.

Un problema relevante que estamos enfrentando con el desarrollo de los nuevos sistemas avanzados de IA es la dificultad cada vez mayor que existe para identificar dónde se ha producido un fallo, cuando una acción realizada con una máquina dotada de inteligencia artificial causa daños a un tercero. Ello está siendo una cuestión

1. Estos temas los hemos abordado en dos estudios previos: Parra y Concha (2021 y 2022).

especialmente crítica desde la particular perspectiva de las interacciones de la IA con el derecho de daños. En sistemas basados en aprendizaje automático y redes neuronales profundas, se está observando que los modelos pueden llegar a una respuesta correcta de manera rápida y precisa, pero se opacan a la hora de ofrecer información de cómo llegaron a ese resultado. Ello pone en tensión las exigencias tradicionales respecto de la acreditación del nexo causal y la culpa sobre las que se ha construido históricamente la noción de responsabilidad civil.

Desde esta perspectiva, profundizaremos en la necesidad de reducir la opacidad técnica de los sistemas avanzados de IA, para transitar hacia una inteligencia artificial explicable (*explainable artificial intelligence* o XAI) que dé respuestas tanto al interés de una gobernanza de la IA —que nos lleve a una inteligencia artificial fiable— como a la exigencia de implementar un régimen de responsabilidad civil que no limite el desarrollo y el uso de sistemas de inteligencia artificial cada vez más complejos. Metodológicamente, el artículo se basa en un análisis dogmático y de política legislativa de los principales instrumentos internacionales —especialmente de la Unión Europea—, complementado con literatura sobre gobernanza de la IA y la XAI. Ello permite, por un lado, generar propuestas para avanzar hacia un marco normativo y de gobernanza en torno a la IA fiable y, por otro, evaluar críticamente la forma en que la explicabilidad —o su ausencia— incide en la configuración actual y futura de la responsabilidad civil por daños relacionados con el uso de sistemas de inteligencia artificial.

La gobernanza de la inteligencia artificial

La gobernanza de la inteligencia artificial está siendo uno de los desafíos más complejos que enfrentan los ordenamientos jurídicos. Como se ha esbozado en la introducción, la rápida expansión de la IA —en particular de los sistemas avanzados basados en aprendizaje automático y modelos neuronales profundos— ha tensionado los marcos institucionales existentes, poniendo en evidencia la necesidad de un enfoque más robusto, multidimensional y global de su regulación. Asimismo, hay un rasgo inherente a los sistemas avanzados de IA: su opacidad estructural —no poder entregar una explicación de cómo llegan a las respuestas o soluciones que ofrecen a un problema planteado— está generando una tensión profunda con los principios constitucionales y jurídicos tradicionales, desde la motivación de las decisiones hasta la rendición de cuentas.

En este sentido, señala Robles (2020: 2) que «el desarrollo de una tecnología como la IA que aspira a compararse con la inteligencia humana, o incluso a mejorarla o superarla, constituye posiblemente el mayor desafío científico conocido en la historia de la humanidad». El uso cada vez mayor de la inteligencia artificial en la automatización de procesos sociales y económicos que eran realizados, históricamente, por seres

humanos ha dado lugar a la necesidad de regular su aplicación en sociedades crecientemente digitalizadas. Su naturaleza de tecnología de uso general y su caracterización como una caja negra donde se llevan a cabo procesos que entregan resultados o se toman decisiones sin mostrar o explicar cómo lo ha hecho han convencido a la comunidad internacional de establecer un marco global, holístico y multidimensional para la IA (Antonov, 2022: 42).

Uno de los recientes esfuerzos de la comunidad internacional por dotarse de un marco internacional en IA lo constituye la Resolución 78/265 de la Asamblea General de Naciones Unidas, de 21 de marzo de 2024, que exhorta a los países a adoptar un enfoque común de gobernanza de la IA para aprovechar las oportunidades de sistemas seguros y fiables de IA para el desarrollo sostenible, donde se insiste en la necesidad de sistemas seguros, fiables y de confianza. Asimismo, en el ámbito de la Unión Europea, en 2024 se adoptó el Convenio Marco del Consejo de Europa sobre IA y Derechos Humanos, que constituye el primer acuerdo internacional jurídicamente vinculante en el ámbito de la inteligencia artificial, que establece obligaciones para asegurar que cualquier actividad relacionada con la IA sea compatible con los derechos humanos, los valores democráticos y el Estado de derecho (Council of Europe, 2024).

No obstante, este incipiente marco jurídico internacional, aunque relevante, sigue sin ofrecer un modelo de gobernanza global coherente. La emergente regulación que se está llevando a cabo de las diversas cuestiones que plantea el uso de IA todavía está fragmentada, se confunde con principios éticos y ha tenido una escasa efectividad a nivel global (Robles, 2020: 5). Todo ello, en un contexto de tensiones políticas e intereses económicos contrapuestos que dificultan la necesaria tarea de armonización normativa. En este marco de gobernanza débil, la IA sigue evolucionando de forma acelerada. Uno de los avances técnicos más relevantes es el desarrollo de sistemas expertos, programas informáticos que, desde un conocimiento base entregado por especialistas a través de la codificación del conocimiento, resuelven problemas intentando reproducir procesos del pensamiento humano (Parra y Concha, 2021: 3).

Existen dos tipos de sistemas expertos: los modelos de procesamiento simbólico y los conexionistas.[2] El primero, predominante en la primera fase de la evolución de la

2. Los modelos de procesamiento simbólico o basados en conocimiento están fundados en el razonamiento simbólico. Es decir, simulan las capacidades cognitivas de los seres humanos mediante el procesamiento de fórmulas sintácticas que reflejan la estructura del lenguaje. De acuerdo con este enfoque, tanto el ser humano como el ordenador procesan la información que recogen desde el entorno, a partir de la manipulación de símbolos discretos. En ese sentido, ambos son sistemas que usan la lógica de primer orden para procesar secuencialmente la información que reciben del ambiente (Amoruso, Bruno y Dominino, 2007: 338). Los modelos conexionistas o neuronales tratan de resolver problemas no algorítmicos a partir de la experiencia almacenada como conocimiento. Es decir, buscan entender cómo funciona el cerebro humano y replicar su comportamiento. En un modelo conexionista la información

IA, es un enfoque que representa el conocimiento mediante símbolos y reglas explícitas, operando a través de manipulaciones lógicas y estructuras formales para emular procesos de razonamiento humano (Russell y Norvig, 2002). Por otro lado, con el desarrollo del modelo conexionista de redes neuronales artificiales, las operaciones de la IA han comenzado a ser cada vez menos explicables. De acuerdo con el proyecto Construcción Internacional de Capacidades para la Evaluación y Gobernanza de la Biología Sintética: «Los actuales sistemas de aprendizaje automático, que usan redes neuronales, pueden llegar a la "respuesta correcta" a un problema con creciente precisión, pero no son capaces de ofrecer una explicación de cómo llegaron a ella».[3]

El avance de IA avanzada, con cada vez mayores niveles de autonomía, pone en evidencia la debilidad del marco normativo y de gobernanza actual para responder a los desafíos que esta tecnología nos está planteando. En este sentido, Robles (2020: 4) señala que las diversas iniciativas, en los ámbitos internacional, regional y estatal, han comenzado a mostrar los contornos de un esquema de gobernanza de la IA, que suelen partir de dos presupuestos erróneos derivados del abrumador desarrollo tecnológico relacionado con la inteligencia artificial: por una parte, una visión antropomórfica de la misma, que se suele plasmar en las normas jurídicas, y, por otra, una concepción unitaria u homogénea de la IA, que tampoco es real.

Tal como expone Dafoe (2018: 5), existen probabilidades de que la IA llegue, dentro de este siglo, a dar respuestas inmediatas a gran parte de las tareas importantes con unas prestaciones «sobrehumanas». En este contexto, surgirían capacidades de la IA que transformen drásticamente el bienestar, la riqueza o el poder, en dimensiones comparables a las revoluciones industrial y nuclear. Por otra parte, de forma mayoritaria, el estudio de la inteligencia artificial se ha ido decantando por el análisis interdisciplinario, atendiendo a su alcance global y su naturaleza difusa (Solarczyk, 2017). Ello ha tenido, según Robles (2020: 4-5), un efecto pernicioso desde la perspectiva del derecho, referido a que el análisis de los diversos aspectos relacionados con esta tecnología está realizándose principalmente desde la técnica y la ética, bajo el formato de principios éticos o estándares técnicos, reduciendo el papel que debe jugar el derecho.

---

es procesada de manera paralela por medio de un conjunto de unidades que interactúan entre sí de un modo similar a las neuronas del cerebro humano (Amoruso, Bruno y Dominino, 2007: 338). Siguiendo los modelos conexionistas, se construyen máquinas inteligentes que simulan el cerebro humano, usando neuronas artificiales. De esta forma, por medio de un ordenador, se modela el cerebro, imitando la arquitectura y funciones de las conexiones neuronales (Martínez, 2012: 832).

3. Proyecto Construcción Internacional de Capacidades para la Evaluación y Gobernanza de la Biología Sintética, «Biología sintética y biosíntesis habilitada por IA: Implicaciones para la biodiversidad y la subsistencia campesina. Informe para los delegados del Convenio sobre Diversidad Biológica», Building International Capacity on Synthetic Biology Assessment and Governance Project, 2018, página 11, disponible en https://tipg.link/mb6w.

Uno de los problemas que ello ha traído consigo es, precisamente, los escasos avances que se han producido en el desarrollo de un marco general de gobernanza de la inteligencia artificial. Si bien, como hemos señalado, han existido diversas iniciativas, los resultados obtenidos no permiten hablar todavía de un estatuto universal para la IA, más allá de la identificación de un conjunto de principios éticos que promueven el desarrollo de una inteligencia artificial respetuosa de los valores democráticos, cuestión en la que ha venido trabajando en los últimos años intensamente la Unión Europea (Solar, 2020; Parra y Concha, 2022). Expresa Dafoe (2018: 5-6) que, cuando hablamos de gobernanza de la inteligencia artificial, nos preguntamos cómo la humanidad debe afrontar de mejor forma la transición a los sistemas avanzados de inteligencia artificial desde la dimensión política, económica, militar, gubernamental y ética. De esta forma, uno de los obstáculos para avanzar hacia una gobernanza de la IA es la exigencia de encontrar consensos sobre cómo debemos prepararnos para este escenario, integrando todas las implicaciones de alto riesgo de la IA avanzada.

Sabemos que la IA tiene unas ventajas potenciales en todos los campos de nuestra vida —salud, calidad de vida, riqueza, sostenibilidad, ciencia, reducción de accidentes, integración social, educación, democracia y justicia (Comisión Europea, 2021a: 1; Cantarini, 2023: 122)—. Pero también, que existen amenazas, en algunos casos muy elevadas, y la gobernanza de la inteligencia artificial nos prepara para los riesgos que comienzan a ser cada vez más intensos a medida que se logran mayores avances en IA. En ese sentido, la gestión de la escala y la velocidad en que estamos adoptando la inteligencia artificial y sus riesgos relacionados está siendo una tarea cada vez más central para los gobiernos. No obstante, en muchos casos, los beneficiarios de estas tecnologías no asumen los costos de dichos riesgos, que son transferidos a la sociedad o a los gobiernos (Taeihagh, 2021: 138). Desde esta perspectiva, Dafoe (2018: 7-8) agrupa en cuatro las principales fuentes de riesgo grave derivadas de la IA avanzada:

- La ayuda a la consolidación de los totalitarismos: la manipulación de los procesos democráticos, como las elecciones a través de las redes sociales, es un claro ejemplo de ello.[4]
- Guerra (nuclear) preventiva, involuntaria o incontrolable entre grandes potencias: la IA podría dar lugar a ventajas extremas de primer ataque o nuevas capacidades destructivas de gran impacto que podrían tentar a una gran potencia a iniciar una guerra.
- Construcción de sistemas IA ampliamente sobrehumanos que no estén totalmente alineados con los valores humanos, dando lugar a la extinción humana

4. Para profundizar en esta idea, véanse Santana y Huerta (2019); y Gorodnichenko, Pham y Talavera (2021).

o la pérdida permanente de valores: en este sentido, los principales esfuerzos de la Unión Europea en impulsar una gobernanza de la IA se han centrado en la elaboración de principios éticos para esta tecnología, relacionados con los valores democráticos y el respeto a los derechos humanos.

- Erosión sistemática de los valores humanos a causa de la competencia: incluso si los tres riesgos anteriores no se concretan, nos enfrentamos a una situación compleja si las empresas que producen IA avanzada se ven en la disyuntiva de perseguir los valores humanos o adaptarse a la competencia.

Sobre estos riesgos, especialmente difíciles de gestionar, es donde debe incidir la gobernanza de la IA, con instituciones adecuadas y un marco normativo general que establezca los consensos de la comunidad internacional. En este sentido, quien más ha trabajado en el impulso de un marco de gobernanza de la inteligencia artificial es la Unión Europea, configurando un marco general europeo para la inteligencia artificial, con la aspiración de que sirva de base para un acuerdo global en la materia.

En cuanto a las dificultades de avanzar hacia un marco general de gobernanza para la IA, Robles (2020: 10-12) señala que, en parte, se debe a que el debate ha sido dirigido desde la tecnología y la ética. Ello lleva a dos problemas principales. En el caso del discurso liderado desde la técnica, el debate sobre la gobernanza de la inteligencia artificial está siendo sustituido por la gobernanza a través de la IA. De esta forma, estamos pasando a un modelo de gobernanza mediante el algoritmo, que incluye esquemas de regulación entregados a la inteligencia artificial en los diversos ámbitos de nuestra vida, sin haber resuelto previamente los problemas de gobernanza de la IA, especialmente los referidos a los riesgos de la IA avanzada. En cuanto al debate ético sobre la inteligencia artificial, este ha desplazado al jurídico o se ha confundido con el mismo. Ello ha tenido dos consecuencias principales: por una parte, la búsqueda de unos valores homogéneos que sean parte de una ética universal ha limitado los avances en la construcción de un marco jurídico general robusto; y, por otra parte, ha llevado a reducir el ámbito normativo a la identificación de dichos valores éticos con principios jurídicos, sin que se aumenten los esfuerzos por un desarrollo normativo e institucional suficiente para hablar de un marco regulatorio propio de la inteligencia artificial.

Debemos tener en consideración que, para avanzar en la gobernanza de la IA, la comunidad internacional debe evaluar la eficacia de los enfoques tradicionales de gobernanza, como el uso de determinados esquemas regulatorios, impuestos y subvenciones ante una tecnología que está sometida a cambios constantes y donde existe una importante debilidad en el acceso a la información. Desde esta perspectiva, se comienza a hablar de modelos de gobernanza emergentes y nuevas estrategias, como

la gobernanza adaptativa o híbrida[5] y el refuerzo del rol de la autorregulación y los estándares técnicos para solventar las debilidades de los escasos progresos logrados hasta ahora en la materia (Taeihagh, 2021: 138-139; Aneja, 2021: 2-3).

Un desafío que todo ello plantea, relacionado con el objeto de estudio, es si los avances en gobernanza de la IA —que tienen como uno de sus puntos neurálgicos la explicabilidad en los discursos de esta gobernanza— se traducen en obligaciones jurídicas efectivas o incentivos estructurales para adoptar un modelo globalmente reconocido, o si en realidad nos están llevando a aprender a convivir con la opacidad técnica de los sistemas de IA avanzada. Esta misma idea, como veremos más adelante, es aplicable a los regímenes de responsabilidad civil.

La propuesta de la Unión Europea en materia de gobernanza de la inteligencia artificial

Los avances de la Unión Europea en materia de gobernanza de la inteligencia artificial se construyen sobre el objetivo fundamental de avanzar hacia una IA fiable y centrada en el ser humano. La revisión de sus últimas propuestas en la materia es especialmente relevante, porque, como hemos dicho, la Unión Europea no solo busca establecer un marco general europeo de gobernanza de la IA para Europa, sino que también aspira a que este sea la referencia para un futuro esquema de gobernanza global para la inteligencia artificial (Comisión Europea, 2018: apartado 3.3).

La estrategia europea sobre IA está establecida en el Reglamento (UE) 2024/1689, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Parlamento Europeo y Consejo, 2024a). En su considerando 1 establece que su finalidad es «mejorar el funcionamiento del mercado interior uniforme, en particular, para el desarrollo, la introducción en el mercado, la puesta en servicio y la utilización de sistemas de IA en la Unión, de conformidad con los valores de la Unión»: i) a fin de promover la adopción de una inteligencia artificial centrada en el ser humano y fiable; ii) garantizando al mismo tiempo un elevado nivel de protección de la salud, la seguridad y los derechos fundamentales consagrados en la Carta de los Derechos Fundamentales de la UE, incluidos la democracia, el Estado de derecho y la protección del medio ambiente; iii) protegiendo a las personas frente a los efectos perjudiciales de los sistemas de IA en la Unión; y iv) brindando apoyo a la innovación.

En una primera aproximación, debe destacarse que la Unión Europea articula su gobernanza de la IA conciliando dos dimensiones complementarias: i) la construcción del mercado interior, que exige seguridad jurídica y libre circulación de

5. Díez-Gutiérrez (2021: 113) se refiere a la gobernanza híbrida para hablar de esquemas mixtos de gobernanza, caracterizados por la colaboración público-privada.

productos (y datos); y ii) la protección de los derechos fundamentales y los valores democráticos. De esta forma, el establecimiento de un marco general europeo (y global) de gobernanza para la inteligencia artificial debe armonizarse con la libertad para diseñar, crear, comercializar y utilizar sistemas de IA. Por ello, el reglamento se conecta con la estrategia digital europea y con la necesidad de fortalecer el mercado único digital, que incluye la libre circulación de datos no personales (considerando 148). En este sentido, señalaba la comunicación que contenía la propuesta de reglamento de 2018 que este marco general debe avanzar hacia un esquema que permita el desarrollo de un mercado interior de la IA que garantice un «entorno jurídico predecible», que permita a los ciudadanos y las empresas confiar en la tecnología con que interactúan (Comisión Europea, 2018).

Para avanzar en esta confianza de la ciudadanía y las empresas en la IA, la Comisión Europea, en su propuesta de 2018, señalaba que era necesario «comprender de qué modo funciona la tecnología, de ahí la importancia de la investigación sobre la explicabilidad de los sistemas de IA» (apartado 3.3). De esta forma, el marco normativo general y de gobernabilidad europeo que se cristalizó con el reglamento de 2024 promueve una gobernanza que guíe el desarrollo de una IA fiable y centrada en el ser humano.

Esta dualidad del enfoque europeo sobre inteligencia artificial, que concilia el desarrollo de un mercado interior para la IA que permite aprovechar todos los beneficios de esta tecnología con la necesidad de garantizar un marco jurídico predecible y fiable, es recogida por el *Libro blanco sobre la IA*, de 2020. En él se desarrolla el planteamiento que ofrecía la Unión Europea, primero en la Estrategia Inteligencia Artificial para Europa (Comisión Europea, 2018), centrada en la necesidad de «aprovechar las oportunidades que ofrece la inteligencia artificial y abordar los retos que presenta, la Unión Europea debe actuar conjuntamente y determinar de qué manera, a partir de los valores europeos, promoverá su desarrollo y adopción» (Comisión Europea, 2020: 2), que sería recogida y transformada en norma jurídica posteriormente con el AI Act, de 2024 .

En este sentido, el *Libro blanco sobre la IA* reafirmó la importancia de los datos. El desarrollo, económico, ambiental y social de la Unión Europea se apoya cada vez más en los valores creados por los datos, y la IA es una parte fundamental de la denominada economía de datos. Desde esta perspectiva, la inteligencia artificial «es una combinación de tecnologías que agrupa datos, algoritmos y capacidad informativa» (Comisión Europea, 2020: 2). Por tanto, para ser líder en IA, Europa requiere: i) explotar el potencial tecnológico que permita mayores avances en computación; y ii) disponer de una infraestructura digital de calidad, que permita almacenar y manejar la creciente disponibilidad de datos. Así, la Unión Europea podrá convertirse en un líder global de la innovación en la economía de los datos y sus aplicaciones (Comisión Europea, 2020: 2).

Como contrapartida, la Comisión reconocía la necesidad de avanzar previamente en la confianza en el uso de la IA, a través de un marco general de gobernanza que se base en sus valores esenciales. De esta forma, la Unión Europea con el AI Act ofrece desarrollar un ecosistema de IA que lleve los beneficios de esta tecnología a los ciudadanos, las empresas y a los servicios de interés público, en un entorno seguro y fiable. Podemos observar, entonces, que, desde la perspectiva de la Unión Europea, el marco jurídico general para una gobernanza de la IA busca ofrecer un ecosistema de inteligencia artificial que permita generar confianza en los ciudadanos y las empresas, para desarrollar y utilizar sistemas de IA en todas las fases posibles de nuestra vida. Para conseguir ambas dimensiones de la gobernanza de la inteligencia artificial, el «Libro blanco» promueve el impulso de, por una parte, un marco político que desarrolle medidas de armonización de los esfuerzos a nivel europeo, regional y nacional. Ello incentivará, en un contexto de colaboración público-privado, la movilización de recursos para conseguir un «ecosistema de excelencia» que permita el desarrollo de una economía basada en la IA. Y, por otra, un marco normativo para la inteligencia artificial en Europa que genere un «ecosistema de confianza». Este debe abarcar temas especialmente sensibles como el respeto y la protección de los derechos fundamentales, el resguardo de los derechos de los consumidores, la seguridad jurídica a las empresas y organismos públicos para que inviertan en innovación en esta tecnología (Comisión Europea, 2020: 3).

Centrándonos en la segunda dimensión referida —ecosistema de confianza—, el «Libro blanco» establece que uno de los principales desafíos del futuro marco normativo general es conciliar las exigencias de protección de los derechos de los ciudadanos con la necesidad de las empresas de contar con un régimen que garantice la seguridad jurídica y, con ello, permitir aumentar la inversión en I+D en IA. La forma de resolver este problema es generar confianza en los ciudadanos, las empresas y el sector público, es decir, avanzar hacia una IA fiable (Comisión Europea, 2020: 11).

La Comisión Europea profundiza en esta idea en su comunicación «Fomentar un planteamiento europeo en materia de IA», de 2021, reconociendo que la ciudadanía europea desconfía del uso de la IA por los riesgos elevados en ciertas materias en las que la legislación existente no da respuesta (derechos fundamentales, seguridad, derechos de los consumidores). Desde esta perspectiva, propone que el marco general de gobernanza de la inteligencia artificial debe estar basada en el riesgo. Por ello, se insiste en que el futuro marco normativo general contemple el doble objetivo de abordar los riesgos asociados a la tecnología IA de una manera proporcionada, al mismo tiempo de promover la inteligencia artificial (Comisión Europea, 2021a: 6-7).

Todos estos planteamientos fueron recogidos en el Reglamento (UE) 2024/168, de inteligencia artificial, que se fundamenta en un enfoque basado en el riesgo, mediante el cual las obligaciones aplicables a un sistema de IA dependen del tipo y nivel de riesgo que dicho sistema puede generar para los derechos fundamentales, la seguridad, la

salud o los valores democráticos. La llamada Ley Europea sobre Inteligencia Artificial clasifica los sistemas de IA en cuatro categorías: i) riesgo inaceptable; ii) los sistemas de alto riesgo; iii) riesgo limitado; y iv) riesgo mínimo.

De esta forma, en primer lugar, el reglamento establece un conjunto de prácticas de IA prohibidas, que conllevan riesgos inaceptables por ser contrarios a los valores de la Unión Europea, como aquellas actividades que involucran la vulneración de derechos fundamentales. El artículo 5.1 contempla ocho supuestos de prácticas de inteligencia artificial prohibidas, entre ellas, el uso de sistemas para explotar vulnerabilidades de personas o colectivos, por razones de edad, discapacidad, situación social o económica concreta, o para evaluar riesgos de criminalidad de personas basándose en su perfil o rasgos.

El reglamento regula de los sistemas de alto riesgo —que son el núcleo de este—, aceptando su introducción al mercado europeo, siempre que cumplan con ciertos requisitos obligatorios, referidos a la gobernanza de datos, documentación técnica, registros para asegurar su trazabilidad, vigilancia humana, transparencia, solidez y ciberseguridad (artículos 9 a 15). Además, estos sistemas deberán ser evaluados *ex ante*. En este sentido, el reglamento opta por un marco general con un grado de intervención reguladora intermedia, que se aplique únicamente a los sistemas de IA de alto riesgo. Finalmente, los sistemas de IA no clasificados de alto riesgo —es decir, de riesgo limitado o mínimo— se podrán regir por códigos de conducta, aunque la Unión Europea promueve que se sometan voluntariamente a las reglas para inteligencia artificial de alto riesgo.

Por otra parte, el AI Act establece una estructura institucional multinivel para la gobernanza de la IA. En el ámbito comunitario, se establece un Comité Europeo de IA encargado de vigilar el cumplimiento del reglamento, de la coordinación entre Estados miembros y la supervisión a los mismos. A nivel nacional, los Estados miembros deberán designar autoridades nacionales que supervisarán la aplicación del reglamento. Además, se establece una base de datos europea donde se incorporan los sistemas de IA de alto riesgo independientes, que incluirá, entre otros, los datos de los proveedores de estos sistemas. Finalmente, destacamos que se entregará atribuciones a las autoridades de vigilancia del mercado para actuar sobre estos sistemas de IA de alto riesgo una vez introducidos en estos (artículos 10 y siguientes).

Una de las innovaciones más relevantes del AI Act es la regulación específica del modelo de IA de uso general, que es aquel «entrenado con un gran volumen de datos utilizando autosupervisión a gran escala, que presenta un grado considerable de generalidad y es capaz de realizar de manera competente una gran variedad de tareas distintas» (artículo 6, número 63), incluidos los modelos fundacionales y los sistemas generativos basados en redes profundas. Los modelos de IA de uso general quedan sujetos a obligaciones adicionales, entre ellas (artículos 52 a 55): i) documentación técnica amplia y accesible; ii) divulgación del uso de datos protegidos por derechos

de autor; iii) políticas para prevenir usos indebidos; y iv) evaluación y mitigación de riesgos sistémicos en el caso de modelos de alto impacto. Esta categoría refleja que el modelo europeo reconoce la opacidad estructural y la dificultad de explicar decisiones generadas por modelos de gran escala. Por ello, el reglamento desplaza el énfasis desde la explicabilidad algorítmica hacia la trazabilidad, auditabilidad y documentación, es decir, la explicabilidad estructural que hemos mencionado antes.

Si bien, el AI Act supone un gran avance en la regulación y la gobernanza de la inteligencia artificial, puede señalarse que se mantienen tensiones en los ámbitos de la sobrerregulación y su desincentivo a la innovación, la dependencia de estándares técnicos privados, la diversidad regulatoria nacional y el reconocimiento implícito a las limitaciones de mayores avances en la explicabilidad (Ebers y otros, 2021).[6] De esta forma, la Unión Europea busca conciliar los dos enfoques que están detrás del marco general de gobernanza: uno con enfoque en los derechos fundamentales y el otro centrado en el riesgo de la innovación (Antonov, 2022: 42). Este modelo dual es la propuesta de la Unión Europea para lograr una IA fiable.

Desde la perspectiva de la explicabilidad, aunque el AI Act no introduce un derecho subjetivo general a obtener explicaciones en el marco del uso de la IA, sí incorpora obligaciones que buscan reducir la opacidad de la IA. En los sistemas de alto riesgo, por ejemplo, exige que su diseño y desarrollo garantice un nivel adecuado de transparencia (artículo 13), la provisión de instrucciones claras para comprender el funcionamiento del sistema, la capacidad de generar registros que permitan realizar auditorías y la obligación de garantizar una supervisión humana efectiva (artículos 14 y 15). Todas estas obligaciones reflejan un cierto giro institucional en la Unión Europea hacia lo que podemos denominar como una explicabilidad estructural, que apunta no tanto a explicar de manera concreta cada resultado individual, sino en garantizar condiciones para la trazabilidad, la responsabilidad y la auditabilidad del sistema.

Hacia una inteligencia artificial fiable

El uso de la inteligencia artificial «genera una serie de riesgos elevados específicos a los que la legislación existente no da respuesta» (Comisión Europea, 2021a, apartado 4). De manera general, los países han ido adaptando sus regímenes sectoriales para ir enfrentando los diversos desafíos que se van presentando con el uso cada vez más intensivo de la IA en múltiples actividades. Así, por ejemplo, en el ámbito de la propiedad de los sistemas de inteligencia artificial, los países han ido considerando a la IA como una obra producida por la actividad creativa y, en consecuencia, ha sido protegida como tal por los regímenes de propiedad intelectual. Por otra parte, el software

6. Anthony Rutkowski, «The EU AI Act: A critical assessment», *CircleID*, 28 de junio de 2023, disponible en https://tipg.link/mb9x.

lo ha sido a través de los derechos de autor y del régimen de patentes. No obstante, existen ciertos problemas con estos esquemas; por ejemplo, algunos sistemas de IA están incorporados a dispositivos físicos como los robots, que se consideran productos y, desde una perspectiva del derecho de daños, son regulados por los regímenes referidos a los productos defectuosos (Solarczyk, 2017: 57-58).

Este ejemplo sirve para exhibir la dificultad que está teniendo la respuesta a los problemas que se presentan con el uso de los sistemas de IA, por parte de los regímenes jurídicos vigentes. Lo mismo sucede en el ámbito de los derechos laborales, los esquemas contractuales o de los derechos relacionados con la privacidad de la persona, entre otros. Por ello, a medida que nuestra interacción con la tecnología de IA sea más compleja, mayores dificultades tendremos para resolver las controversias surgidas con los sistemas de IA cada vez más autónomos y que manejan una alta cantidad de datos.

En este sentido, desde hace algunos años se ha venido expresando la necesidad de avanzar hacia regímenes jurídicos específicos para la IA (AI HLEG, 2019; Santos, 2017). Las controversias de naturaleza jurídica que deben ser resueltas son cada vez más complejas y los estatutos tradicionales comienzan a verse sobrepasados. Por tanto, el reto a que nos enfrentamos desde una perspectiva jurídica es doble: en el ámbito sectorial, se debe llevar a cabo un proceso de adaptación de los diversos regímenes jurídicos que se relacionan con la inteligencia artificial —laboral, ambiental, contractual, tránsito, y seguridad vial responsabilidad civil, ciberseguridad, entre otros—; y, de forma general, es necesario avanzar hacia un marco normativo y de gobernanza, de carácter global, que regule los aspectos básicos, como hace la propuesta del Reglamento de la Comisión Europea de 2021: prohibición de las prácticas de IA contrarias a los derechos fundamentales y los principios democráticos; la diferenciación entre sistemas de alto riesgo y otros sistemas de IA para en el establecimientos de requisitos para su creación, comercialización y operación; reglas básicas para la transparencia; propuesta de un modelo de gobernanza de la IA, con una institucionalidad especial; y otras cuestiones referidas a la innovación, entre otras.

El concepto de inteligencia artificial fiable como objetivo del marco normativo es aplicable a ambas dimensiones, es decir, tanto para la adaptación del ordenamiento sectorial como para el nuevo marco general normativo y de gobernanza de la IA. De esta forma, si bien centraremos el análisis en este último aspecto, también es aplicable a los diversos regímenes jurídicos específicos que deben adaptarse a los desafíos que propone la regulación de la IA. Como hemos señalado, la Unión Europea ha puesto énfasis en que el nuevo marco general para la inteligencia artificial debe conducir hacia una IA fiable centrada en el ser humano. Creemos que este es un planteamiento adecuado, por cuanto concilia la necesidad de tener unas reglas universales sobre aspectos centrales de la IA, con la posibilidad de aprovechar todos los beneficios que podemos obtener con el desarrollo de esta tecnología.

Desde esta perspectiva, la estrategia de la Unión Europea en IA se articula sobre la base de dos dimensiones esenciales: el impulso al mercado interior para esta tecnología, incluido el mercado único digital europeo; y el respeto a los valores de la Unión Europea, encabezados por los derechos fundamentales y los principios democráticos (Parlamento Europeo, 2022). En este sentido, una IA fiable y centrada en la persona es el estándar normativo que concilia ambas dimensiones y debe constituir el punto de partida de un marco normativo y de gobernanza global para la inteligencia artificial.

El Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial, creado por la Comisión Europea, elaboró en 2019 unas directrices que establecen cuatro principios éticos que sirven de fundamento de una IA fiable, ética y robusta, basada en los derechos humanos: el respeto de la autonomía humana, la prevención del daño, la equidad y la explicabilidad. La fiabilidad ética, entonces, se refiere a los ámbitos frontera de la ética/derecho de la IA, y la robustez se refiere a la fiabilidad técnica, que se contiene especialmente en el principio de la explicabilidad que se analizará en uno de los siguientes apartados. El enfoque en derechos fundamentales es, como hemos venido señalando, el fundamento de una IA fiable. Forman parte de este enfoque, que es un límite a las actividades que se pueden realizar con esta tecnología: la dignidad humana y la libertad individual; la democracia, la justicia y el Estado de derecho; la igualdad, no discriminación y solidaridad; y los derechos de los ciudadanos, especialmente la participación y el derecho a la buena administración (AI HLEG, 2019: 12-13). Los cuatro principios que orientan la implementación de una IA fiable deben seguir este enfoque en los derechos fundamentales.

El principio de respeto a la autonomía humana obliga a que, en la interacción con la tecnología IA, debemos conservar una autonomía plena y efectiva sobre nosotros mismos, sin que la inteligencia artificial pueda subordinarnos, engañarnos, coaccionarnos, dirigirnos, condicionarnos o manipularnos de forma injustificada. Por el contrario, la IA debe ser diseñada para aumentar, potenciar y complementar nuestras aptitudes cognitivas, culturales y sociales. Para ello, es necesario que se siga un diseño centrado en las personas.

En virtud del principio de prevención del daño, los sistemas de IA no deben causar daños o agravar los que ya existen. Tampoco pueden perjudicar de cualquier manera a los seres humanos, incluidas su dignidad e integridad física y mental. Para lograrlo, todos los sistemas de IA deben ser seguros y robustos técnicamente. El principio de equidad obliga a que el desarrollo, despliegue y utilización de los sistemas de IA sean equitativos. Ello significa que, al menos, haya una distribución justa y equitativa de los beneficios y costes, y evitar la discriminación o estigmatización. Por último, el principio de explicabilidad determina unos procesos transparentes, en que se explique clara y abiertamente las capacidades y finalidad de los sistemas de inteligencia artificial. Asimismo, las decisiones que adopte la IA deben ser explicables a quienes sean afectados, directa o indirectamente por ellas.

Además de estos principios orientadores, en las directrices de 2019 se contemplan siete requisitos clave que deben cumplir los sistemas IA para ser considerados fiables: i) que sean sometidos permanentemente a la acción y supervisión humanas; ii) solidez técnica y seguridad de esta tecnología; iii) gestión de la privacidad y de los datos; iv) transparencia; v) diversidad, no discriminación y equidad; vi) bienestar social y ambiental; y vii) rendición de cuentas (AI HLEG, 2019: 10).

El Parlamento Europeo (2020: apartado 47) señala que para lograr una inteligencia artificial fiable debe seguirse un enfoque del diseño basado en los valores de la Unión Europea y en los principios y requisitos antes enumerados. Solo de esta forma se crearán las condiciones para una aceptación social amplia de la IA y sus tecnologías conexas. Este enfoque europeo se inserta en una tendencia global que avanza hacia principios compartidos sobre IA fiable, presentes en los Principios de IA de la Organización para la Cooperación y el Desarrollo Económicos, adoptados en mayo de 2019 y actualizados en 2024 o en la Recomendación sobre la Ética de la Inteligencia Artificial de la Unesco en 2021.

Desde esta perspectiva, la inteligencia artificial fiable opera como un objetivo que debe guiar a los diversos esquemas normativos que regulan los diversos aspectos jurídicos de la inteligencia artificial, y también como supuesto para generar un entorno favorable para el desarrollo de la IA avanzada. No obstante, la noción de IA fiable sigue siendo criticada por su indeterminación conceptual y por la ausencia de un vínculo jurídico directo entre los principios éticos y las obligaciones legales efectivas.

La inteligencia artificial explicable

Como suele suceder con todo lo referido a IA, no existe consenso sobre una definición de la inteligencia artificial explicable. Existen diversos conceptos según el campo de estudio al que recurramos. Desde un punto de vista técnico, la XAI es un subcampo de la inteligencia artificial, cuyo objetivo es entregar explicaciones de las predicciones, decisiones y recomendaciones de los sistemas inteligentes. En los últimos años, ha tenido un importante desarrollo, debido al énfasis que se ha puesto a la explicabilidad de la IA en el ámbito regulatorio (Ridley, 2022: 1).

Desde una dimensión social, política y jurídica, la XAI es un principio ético que orienta al objetivo de lograr una IA fiable y centrada en la persona. En este sentido, las directrices éticas para una IA fiable de 2019 establecen que la explicabilidad es uno de los «principios éticos, arraigados en los derechos que deben cumplirse para garantizar que los sistemas de IA se desarrollen, desplieguen y utilicen de manera fiable» (AI HLEG, 2019: 14).

El desarrollo de esta técnica tiene lugar en el marco de la evolución que está teniendo el aprendizaje automático, que a su vez es una rama de la IA. De acuerdo con la Defense Advanced Research Project Agency, el objetivo de la XAI es desarrollar

una serie de técnicas de aprendizaje automático, nuevas o modificadas, que generen modelos explicables combinados con técnicas de explicación eficaces que permitan a los usuarios finales comprender, confiar y gestionar de manera segura la nueva generación de sistemas de IA.[7]

En su origen, la inteligencia artificial explicable surgió como una preocupación propia de la informática, es decir, en el ámbito técnico. No obstante, Ridley (2022: 2) señala que con el reglamento de la Unión Europea sobre protección de datos, de 2018, esto pasó a ser una cuestión esencial y urgente de política pública. Desde esta perspectiva, la transparencia como derecho relacionado con el tratamiento de datos incluido en el reglamento involucra a los derechos de revisión, oposición e impugnación de las decisiones automatizadas (Cantarini, 2023: 125).

La explicabilidad busca que los sistemas de IA realicen procesos transparentes, comunicando abiertamente sus capacidades, así como la finalidad de estos. Un elemento central de este principio es que las decisiones deben poder ser explicadas a las personas que puedan verse afectadas por ellas, sea directa o indirectamente (Raposo, 2024). Desde una perspectiva jurídica, la XAI resulta necesaria para:

- Generar confianza, transparencia y comprensión de la IA en los usuarios finales.
- Garantizar el cumplimiento de la regulación sobre IA.
- Mitigar el riesgo que se produce con el uso de la tecnología IA.
- Generar modelos responsables, fiables y sólidos.
- Reducir las posibilidades de un uso de la IA contrario a los derechos fundamentales y los principios democráticos.
- Disminuir el riesgo de sesgo o error.
- Validar modelos de IA y las explicaciones que dan los sistemas de IA (Ridley, 2022: 2).

No obstante, AI HLEG (2019: 16) reconoce que «no siempre resulta posible explicar por qué un modelo ha generado un resultado o una decisión particular (ni qué combinación de factores contribuyeron a ello). Esos casos, que se denominan algoritmos de "caja negra", requieren especial atención». En esta situación, deben adoptarse otras medidas que permitan la explicabilidad, como la trazabilidad, la comunicación transparente sobre los sistemas de IA y sus prestaciones, y la auditabilidad.

Existe un cierto consenso, en el ámbito jurídico, sobre dos límites fundamentales a los sistemas con algoritmos de caja negra en su conjunto: la doctrina de los derechos

7. Véase Defense Advanced Research Projects Agency, «XAI: Explainable artificial intelligence», disponible en https://tipg.link/mb5X.

fundamentales y los principios democráticos. Cuando las decisiones de los sistemas de IA avanzada dejan de ser explicables en ámbitos relacionados con la afectación de derechos fundamentales o la vulneración de los principios esenciales de los Estados democráticos, pasan a ser tecnologías prohibidas normativamente.

Pero la explicabilidad, que parece tan clara y lógica cuando hablamos de la IA fiable, es cada vez menos posible técnicamente a medida que se perfeccionan los sistemas de IA avanzada. Este es el punto central de este trabajo, reflexionar sobre el contenido jurídico de un principio ético que obliga a la transparencia en el desarrollo de una tecnología —la IA avanzada—, que por su propia naturaleza tiende a la inexplicabilidad. De hecho, advierte Ridley (2022: 1), la propia tecnología XAI es también a menudo opaca y compleja.

Esta idea es extrapolable al derecho de daños. La propia evolución de la IA nos está llevando a interactuar con máquinas que tienen progresivamente mayores niveles de autonomía, situación que hace cada vez más difícil una respuesta satisfactoria desde los esquemas y regímenes normativos tradicionales. De esta forma, aun cuando tengamos identificado el problema, necesitamos una IA explicable que la haga fiable, porque de esta manera se reducen los problemas de confianza de las personas, empresas y gobiernos. Y, desde una perspectiva más concreta en el derecho de daños, se permite dar respuestas a las distintas situaciones surgidas en torno a la responsabilidad por los perjuicios causados por sistemas de IA. No obstante, sabemos que la tecnología naturalmente está avanzando hacia la inexplicabilidad, incluida aquella que ha sido creada para dar explicabilidad (la XAI).

Por tanto, estamos ante un dilema: necesitamos una IA explicable porque eso la hace fiable, pero la tecnología busca, en esencia, mejorar la capacidad de imitar o rivalizar con la inteligencia humana en la resolución de problemas complejos, y los avances de esta tecnología empujan a una realidad en la que progresivamente vamos siendo sustituidos o superados por los sistemas de IA (Taeihagh, 2021: 137-138). Si la tecnología es más avanzada en cuanto más logre replicar el funcionamiento de la mente humana, más inexplicable será para nosotros.

En el mismo sentido, Vale, El-Sharif y Ali (2022: 815) señalan que la precisión predictiva del aprendizaje automático es cada vez mayor. Ello ha permitido que sea utilizada para adoptar decisiones de alto riesgo, tanto en el sector público (por ejemplo, la justicia penal) como en el privado (la medicina o la banca). Para avanzar en mayores niveles de precisión predictiva, los algoritmos se hacen cada vez más complejos y, al mismo tiempo, menos transparentes. De esta forma, la precisión se consigue mediante la complejidad del modelo, y ello se acompaña de menores niveles de explicabilidad.

Desde esta perspectiva, Bathaee (2018: 891) pone en evidencia que el debate sobre la IA se centra en dos cuestiones fundamentales: la necesidad de establecer un marco jurídico general y de gobernanza que aborde las principales problemáticas

globales sobre esta tecnología; y adaptar los diversos regímenes que interactúan con la IA, para resolver los problemas que se van presentando con el uso de la inteligencia artificial y anticiparnos al futuro de esta tecnología. No obstante, señala el autor, se presta escasa atención a la cuestión de si los regímenes jurídicos actuales pueden aplicarse de forma correcta a la inteligencia artificial. Y aquí es donde radica el problema principal que nos ocupa: el derecho que aplicamos para abordar el futuro de la IA se basa en un sistema que se centra en la conducta humana (Bathaee, 2018: 892). En este sentido, Krausová (2017: 57-58) explica que el problema de regular la inteligencia artificial o de crear un marco jurídico especializado radica en la propia naturaleza de la inteligencia artificial.

La inteligencia artificial basada en redes neuronales artificiales replica la inteligencia humana, entendida como capacidad para interpretar y aprender de la información. Dichas redes son cada vez más complejas hasta llegar a los modernos algoritmos de aprendizaje automático, que pueden analizar grandes cantidades de datos y adoptar decisiones sin orientación humana directa. La mayoría de los expertos señalan que las aplicaciones de la IA todavía no pueden alcanzar las capacidades cognitivas humanas (IA débil), limitándose a realizar tareas en dominios específicos, aunque ello ya está siendo suficiente para reemplazar a humanos con cualificación media. No obstante, se está avanzando hacia la llamada inteligencia artificial general o fuerte, comparable o superior a la humana, con capacidad para generalizar en contextos que no se han programado previamente (Taeihagh, 2021: 139-140).

Si bien el derecho hasta ahora ha podido dar respuesta a las operaciones de la IA débil de una manera relativamente satisfactoria, con los algoritmos de aprendizaje automático sobre los que se sustenta la IA fuerte, las instituciones tradicionales del derecho podrían no funcionar, por la naturaleza misma de este tipo de tecnología. El avance de la inteligencia artificial está llevando a computadores que ya no se limitan a ejecutar instrucciones previamente entregadas, sino que pueden llegar a soluciones que ni siquiera somos capaces de percibir (Bathaee, 2018: 891-892).

Estamos desarrollando una tecnología con algoritmos de caja negra que obstruye la explicabilidad que se exige para tener una IA fiable, afectando con ello las ideas tradicionales de transparencia, legalidad, justicia y rendición de cuentas (Antonov, 2022: 52). Este problema repercute tanto en el modelo de gobernanza de la inteligencia artificial como en las respuestas que se aportan desde nuestro sistema jurídico que, en el caso concreto de nuestro estudio, ejemplificamos en el derecho de daños. De esta forma, el diseño de algoritmos más justos, responsables y transparentes es una preocupación que está ocupando el debate de la comunidad internacional en los últimos años. En este sentido, Aneja (2021: 30) señala que, en un reciente estudio sobre el algoritmo Compass, utilizado por los tribunales de Estados Unidos, se determinó que se alcanzarían resultados semejantes a los obtenidos si se emplearan menos de los veinte parámetros de información que se utilizan actualmente. Simplificando

el algoritmo, entonces, lograríamos que este fuese más explicable y, con ello, más fácil de supervisar y entender sus respuestas.

Por otra parte, Cantarini (2023: 125) advierte que el debate de la explicabilidad está orientado a la transparencia, el cumplimiento y la responsabilidad de las decisiones automatizadas. De manera general, la transparencia debe estar presente en el diseño de la tecnología de inteligencia artificial, en la recogida y el tratamiento de datos, y en los algoritmos utilizados por el sistema de IA. De esta forma, cuando hablamos de transparencia se abordan dos cuestiones fundamentales: la accesibilidad y la comprensibilidad. Sin embargo, la exigencia de explicabilidad suele estar centrada en la transparencia del algoritmo (romper el código de fuente), que forma parte solo de la accesibilidad. Ello es insuficiente para predecir el comportamiento de los algoritmos, consiguiendo únicamente explicar el método de aprendizaje automático usado por la IA, pero se mantiene oscura la regla de la decisión tomada, dando lugar a lo que se denomina la falacia de la transparencia.

Un conflicto relevante que se da con la XAI es la protección de la propiedad intelectual y los secretos comerciales. Debe encontrarse el equilibrio entre diseñar y desarrollar tecnologías absolutamente explicables con la necesidad de mantener ocultos procesos que forman parte del secreto comercial.[8] Corresponde al marco regulatorio conciliar ambos intereses (Ridley, 2022: 4). A nivel europeo, el AI Act recoge el principio ético de la XAI adoptando obligaciones legales en este sentido. Así, obliga a que los sistemas de alto riesgo sean diseñados y desarrollados asegurando la transparencia y comunicación de información a los responsables del despliegue (artículo 13); tengan una supervisión humana efectiva (artículo 14); y alcancen un nivel adecuado de precisión, solidez y ciberseguridad que funcionen de manera uniforme en estos sentidos a lo largo de todo su ciclo de vida (artículo 15). Así, podemos apreciar que la XAI emerge no como un atributo pleno de los sistemas avanzados de IA, sino como un estándar jurídico que permite gestionar su opacidad estructural. Este punto será determinante para comprender su función en el régimen de responsabilidad civil. A continuación, llevaremos el análisis de la explicabilidad a un ámbito normativo específico, el derecho de daños.

La responsabilidad civil ante la inteligencia artificial explicable

No profundizaremos aquí sobre los diversos esquemas que se han propuesto en el derecho de daños para resolver la indemnización de los perjuicios causados por un sistema de IA, especialmente en el caso de las tecnologías avanzadas, por no ser obje-

8. No profundizaremos sobre este importantísimo tema, por no ser parte del objeto de estudio. Sin embargo, será parte de futuras investigaciones.

to de este estudio.[9] Solo daremos algunas ideas principales, que permitan desarrollar nuestro análisis. En general, puede decirse que las reglas sobre responsabilidad civil son especialmente abiertas y abstractas, y por tanto flexibles, lo que le ha permitido abordar hasta ahora los distintos problemas que se han venido presentando con el daño causado por sistemas de inteligencia artificial (Parra y Concha, 2021: 12). No obstante, los riesgos que han comenzado a surgir con las nuevas generaciones de IA han comenzado a tensionar los esquemas tradicionales del derecho de daños y las respuestas que este ha de entregar.

La responsabilidad civil tiene su fundamento en el principio *alterum non laedere* («no dañar a otro»). Los diversos sistemas jurídicos han adoptado, como regla general, un régimen subjetivo de responsabilidad, integrado por algunos elementos básicos: el daño, la culpa y el nexo causal. Este régimen se ve complementado por un régimen de responsabilidad objetiva, en materia extracontractual, para aquellos casos en que se desarrollan actividades especialmente riesgosas; o en el ámbito contractual, en las obligaciones de resultado (Morales, 2021: 50).

Con la responsabilidad objetiva «se prescinde en absoluto de la conducta del sujeto, no se mira su culpabilidad, se atiende única y exclusivamente al daño producido, basta que este daño se produzca para que el autor del hecho dañino esté obligado a indemnizar» (Díaz, 2007: 82). De esta forma, para las actividades que introducen un riesgo superior en la sociedad los diversos regímenes nacionales han ido incorporando esquemas de responsabilidad objetiva, en que es suficiente demostrar la relación de causalidad entre el hecho y el daño producido para obtener el derecho al resarcimiento de los daños causados por la actividad riesgosa. Esto sucede, por ejemplo, con el régimen de responsabilidad por daños causados por accidentes nucleares o por la navegación aérea.

En este contexto, algunos sistemas jurídicos han optado por incluir a la inteligencia artificial como parte de aquellas actividades peligrosas que se incluyen dentro del esquema de responsabilidad objetiva, debiendo —en caso de daño causado por un sistema de IA— acreditarse el hecho dañino y el nexo causal para activar los engranajes de la responsabilidad civil, con lo que nace para el agente dañoso la obligación de indemnizar aquellos perjuicios provocados sin que sea necesario probar la culpa o negligencia. Como contrapartida, puede «acreditarse el acaecimiento de una causa extraña para exonerarse de responsabilidad civil» (Narváez, 2019: 224).

Diversos regímenes de responsabilidad objetiva han sido utilizados para resolver problemas particulares en el ámbito del derecho de daños y la IA. Así, por ejemplo, en el caso de los automóviles autónomos, se han dejado de utilizar las reglas de responsabilidad civil del conductor y el propietario, para aplicarse, por una parte, las reglas

9. Sobre esta problemática, véase Parra y Concha (2021 y 2022).

de daños debidos a productos defectuosos;[10] y por otra, la responsabilidad de la Administración por los defectos de la infraestructura vial (Parra y Concha, 2021: 11-12).

La Unión Europea dio un paso más, realizando una propuesta de régimen normativo especial para la IA y los robots autónomos en el ámbito de la responsabilidad civil extracontractual. Desde esta perspectiva, el Parlamento Europeo, a través de la resolución de 17 de febrero de 2017, presentó unas recomendaciones para avanzar hacia una personalidad electrónica de los robots y sugirió a la Comisión Europea presentar propuestas para un nuevo régimen de responsabilidad civil para la IA.[11] No obstante, ante el fuerte rechazo que encontró esta propuesta en las demás instituciones de la Unión Europea, el Parlamento rápidamente dio pie atrás, aprobando en 2020 una nueva resolución, en la que reconocía que la personalidad electrónica de los robots no era necesaria porque, técnicamente, los daños o perjuicios causados por la IA terminan siendo de responsabilidad, directa o indirecta, del que lo ha construido, desplegado o interferido en ellos (Parlamento Europeo, 2020; Parra y Concha, 2022: 17).

De esta forma, hubo acuerdo entre las instituciones europeas en dos puntos esenciales: i) que debía adaptarse el régimen jurídico de responsabilidad civil extracontractual para incorporar a la IA, y; ii) que este nuevo marco jurídico no podía ser tan innovador como propuso el Parlamento Europeo en 2017, sino más bien conservador, que no se saliese de los esquemas tradicionales del derecho de daños. Ello se concretó por la Comisión Europea, en septiembre de 2022, con la presentación de dos propuestas legislativas sobre responsabilidad de la IA, que tuvieron distinta suerte:

Directiva (UE) 2024/2853, sobre responsabilidad por los daños causados por productos defectuosos

Esta nueva directiva sobre responsabilidad por los daños causados por productos defectuosos (Parlamento Europeo y Consejo, 2024b) derogó a la Directiva 85/374, de 1985, y supone una integración explícita de las nuevas tecnologías digitales —incluida la IA— en las reglas sobre responsabilidad por los productos defectuosos. La definición de producto se amplía para incluir, no solo a bienes muebles, sino también a los sistemas de IA y a los diversos bienes habilitados para esta tecnología —software, archivos de fabricación digital o actualizaciones o servicios digitales que afecten a la

10. Esta modalidad de responsabilidad objetiva mira a la condición del producto y no a la conducta del fabricante. En el derecho de la Unión Europea, el sistema objetivo de responsabilidad por productos defectuosos fue incorporado por la Directiva 85/374 de 1985. En Chile, se incorporó por la Ley de Protección al Consumidor, aunque de forma muy atenuada, centrada en el derecho a optar par exigir la reparación, reposición o devolución del precio (Díaz, 2007: 100).

11. Sobre esta idea, véase Parra y Concha (2022).

seguridad del producto—. De este modo, el régimen de responsabilidad objetiva por los daños causados por productos defectuosos se extiende tanto a hardware como a software, y servicios digitales interconectados, lo que permite que a los daños relacionados con una IA defectuosa se encuadren plenamente en este esquema de responsabilidad objetiva, que exime a la víctima de la prueba de la culpa del fabricante (Comisión Europea, 2022b; Gómez, 2022: 5).

De esta forma, la directiva mantiene la necesidad de probar los elementos propios de la responsabilidad objetiva: el carácter defectuoso del producto, el daño sufrido y el nexo causal entre el carácter defectuoso y el daño (artículo 10 de la Directiva 2024/2853). La novedad reside en la incorporación de un conjunto de presunciones del defecto y de causalidad que buscan compensar, al menos parcialmente, la complejidad técnica de los sistemas digitales y de IA. Así, se presume el defecto del producto cuando se cumpla alguna de las condiciones siguientes: i) que el demandado incumpla la obligación de revelar las pruebas pertinentes de que disponga; ii) que el demandante demuestre que el producto no cumple los requisitos de seguridad obligatorios que protegen contra el riesgo del daño producido; y iii) que el reclamante demuestre que el daño ha sido causado por un mal funcionamiento evidente del producto durante su uso razonablemente previsible o en circunstancias ordinarias (artículo 10.2).

En cuanto a la relación de causalidad entre el carácter defectuoso del producto y el daño, se presumirá cuando se haya comprobado que el producto es defectuoso y el daño sea compatible normalmente con el defecto (artículo 10.3). Asimismo, el juez presumirá que un producto es defectuoso o el nexo causal entre el referido carácter defectuoso y el daño, si el demandante afronta dificultades excesivas para probar el defecto o el nexo causal, debido a la complejidad técnica o científica, si este demuestra que: i) es probable que el producto fuera defectuoso; o ii) que hay relación de causalidad entre el carácter defectuoso y el daño (artículo 10.4)

En todos estos supuestos, el demandado puede destruir la presunción aportando prueba en contrario. Y aquí es donde puede vincularse de forma directa la cuestión de la explicabilidad de la IA con el régimen de responsabilidad objetiva. En el primer caso (incumplimiento del deber de revelar las pruebas pertinentes), se aprecia una aproximación a la idea de explicabilidad estructural: si el fabricante, importador o distribuidor proporciona documentación, registros y demás información técnica relevante, puede neutralizar la presunción de defecto. Sin embargo, la directiva no exige que esa información permita reconstruir el razonamiento interno del sistema de IA, sino solo que entregue la prueba suficiente disponible. Por tanto, no se impone un deber de explicabilidad algorítmica, sino un deber reforzado de transparencia y cooperación probatoria (Da Fonseca, Vaz de Sequeira y Barreto Xavier, 2024).

El segundo supuesto (incumplimiento de requisitos de seguridad obligatorios) tampoco genera incentivos directos para avanzar hacia una IA explicable. Basta con acreditar que el producto cumple con las normas técnicas y de seguridad aplicables

para desactivar la presunción de defecto, sin que la opacidad o explicabilidad del sistema de IA sea, en sí misma, un criterio relevante. Pero el tercer supuesto (daño causado por un mal funcionamiento evidente durante el uso normal del producto) resulta especialmente significativo para nuestro análisis. Conforme a esta presunción, si el afectado demuestra que el daño se ha producido en condiciones de uso normal y que el comportamiento del sistema se apartó de lo que cabría esperar de un producto seguro, se presume la existencia del defecto. Desde la lógica de la responsabilidad objetiva, esta solución es coherente: el usuario no tiene obligación de desentrañar la lógica interna del modelo de IA, basta con acreditar el resultado dañoso anómalo. Sin embargo, desde la perspectiva de la explicabilidad, este diseño normativo tiene un efecto ambivalente: por una parte, protege al perjudicado frente a la opacidad técnica; por otra, reduce los incentivos para que el fabricante invierta en XAI, ya que la determinación de la causa concreta del fallo deja de ser jurídicamente relevante o necesaria para que surja la obligación de indemnizar.

Con lo dicho anteriormente, se puede decir que el nuevo régimen de productos defectuosos opera como un sustituto funcional de la explicabilidad en litigios de daños: frente a la dificultad de desentrañar lo que ocurre en la caja negra, se recurre a presunciones probatorias. Ello refuerza la protección de la víctima, pero al mismo tiempo consolida la idea de que el sistema jurídico puede convivir con sistemas altamente opacos sin exigir su apertura o comprensibilidad. El resultado es que la XAI no se configura como un deber autónomo del fabricante en el ámbito de la responsabilidad objetiva, sino como una opción estrategia que puede facilitar o no su defensa procesal.

Como vemos, el régimen de responsabilidad por los daños causados por productos defectuosos presenta algunas debilidades para resolver casos complejos de responsabilidad civil de la IA avanzada. Ello es lógico, porque la Directiva no tiene por objeto crear un nuevo régimen de responsabilidad para la IA, sino que busca dar solución a los problemas que esta tecnología está generando a través del régimen de responsabilidad objetiva por daños generados por productos defectuosos, centrado en el consumidor y pensado para problemáticas no complejas.

La fallida propuesta de directiva sobre responsabilidad en materia de inteligencia artificial

Con la presentación de la propuesta de directiva relativa a la adaptación de las normas de responsabilidad civil extracontractual a la inteligencia artificial (Directiva sobre responsabilidad en materia de IA) (Comisión Europea, 2022a), la Unión Europea buscó implementar un régimen especial sobre IA en el ámbito de la responsabilidad civil extracontractual; aunque resultaba ser una propuesta menos disruptiva que las primeras propuestas del Parlamento Europeo en 2017. Con todo, intentaba dar al ré-

gimen de responsabilidad en materia de IA el enfoque europeo coordinado que promueve la Unión Europea desde 2020 (Comisión Europea, 2022a: 1). Sería aplicable a los casos de responsabilidad civil extracontractual subjetiva en aquellas situaciones en que los daños y perjuicios sean causados por un comportamiento ilícito de los sistemas de IA y no entren en el ámbito de la directiva de responsabilidad por productos defectuosos. En concreto, la propuesta establecía normas comunes sobre: i) el acceso a la información relevante de los sistemas de IA de alto riesgo; y ii) la reducción de la carga de la prueba por los daños y perjuicios causados por sistemas de inteligencia artificial de alto riesgo (artículo 1, apartados 1 y 2).

De esta forma, la Unión Europea entregaba unas normas comunes a los Estados para aquellos casos en que el uso de los sistemas de IA tenga un comportamiento ilícito relacionado con, entre otros, las violaciones a la privacidad o los problemas de sesgo y de seguridad. En este sentido, la propuesta de nueva directiva representaba una simplificación de las reglas para acreditar la culpa de una persona que ha provocado daños con el uso de IA, a través de: i) la incorporación de un conjunto de supuestos de presunción de causalidad; y ii) el derecho de acceso a las pruebas (información relevante) que estén en manos de empresas y proveedores, para los casos en que estén involucrados sistemas de IA de alto riesgo (Comisión Europea, 2022b). Respecto del acceso a las pruebas, se establece el deber de los proveedores o usuarios de exhibir las pruebas pertinentes que tengan en su poder sobre un sistema de IA de alto riesgo del que haya sospechas de que ha causado daños, cuando sea requerido por un demandante potencial (artículo 3.1).

En cuanto a presumir, *iuris tantum,* la relación de causalidad en caso de culpa, se establecía la presunción del nexo de causalidad entre la culpa del demandado y los resultados ocasionados por el sistema de IA o la no producción de resultados por parte del mismo cuando: i) hubiera culpa del demandado en el incumplimiento de un deber de diligencia destinado directamente a proteger frente a los daños producidos; ii) sea razonablemente probable que la culpa haya influido en los resultados producidos por el sistema de IA o en la no producción de resultados generados por el mismo; iii) la información de salida producida por el sistema de IA o la no producción de la misma causó los daños (artículo 4.1). Dicha presunción de relación de causalidad podrá aplicarse a los casos de sistemas de IA de alto riesgo, cuando el proveedor ha incumplido algunos deberes relacionados con la fiabilidad técnica, transparencia, vigilancia efectiva de seres humanos o ciberseguridad. No obstante, no será aplicable cuando el fabricante o proveedor ha cumplido con su deber de entregar toda la prueba y conocimientos especializados disponible, si estos son suficientes para demostrar el nexo causal (artículo 4.4).

La fallida propuesta de nueva directiva constituía un avance en materia de acceso a las pruebas relacionadas con los conocimientos técnicos sobre el funcionamiento de los sistemas de IA. No obstante, desde la perspectiva del deber de generar una IA

explicable, el marco de la Unión Europea sigue siendo conservador, reconduciendo el tema a través de las reglas de transparencia establecida en la Ley de Inteligencia Artificial. Como es sabido, esta propuesta sobre responsabilidad en materia de IA fue retirada en 2025 ante la falta de perspectivas de acuerdo político. Ello no significa que los problemas que pretendía resolver hayan desaparecido; más bien, desplaza la carga de la adaptación hacia los Estados miembros, que deberán ajustar sus regímenes de responsabilidad extracontractual a partir del AI Act y de la nueva directiva de productos defectuosos, sin una armonización específica en materia de culpa y prueba.

En este escenario, el AI Act adquiere un protagonismo indirecto en el derecho de daños. Aunque no regula la responsabilidad civil, impone obligaciones de diseño, gobernanza de datos, documentación, transparencia, registro y supervisión humana, especialmente para los sistemas de alto riesgo. El Reglamento 2024/1689 expresa que «los sistemas de IA de alto riesgo se diseñarán y desarrollarán de un modo que garanticen que funcionan con un nivel de transparencia suficiente para que los usuarios interpreten y usen correctamente su información de salida» (artículo 13.1).

En este sentido, la Ley Europea de IA obliga a los sistemas IA de alto riesgo a adjuntar instrucciones de uso que contengan la información completa, correcta y clara, que sea accesible y comprensible a todos los usuarios. Esta información debe incluir la identidad y datos de contacto del proveedor o su representante autorizado; las características, capacidades y limitaciones del funcionamiento de la IA de alto riesgo; los cambios en el sistema de IA de alto riesgo y su funcionamiento predeterminado; las medidas de vigilancia humana previstas; y su vida útil, medidas de mantención y cuidado necesarias para su correcto funcionamiento. Estos deberes configuran un estándar de diligencia técnica que puede ser utilizado por los tribunales nacionales como parámetro de culpa en litigios de responsabilidad subjetiva: el incumplimiento de las obligaciones del AI Act podrá ser interpretado como infracción del deber de cuidado, facilitando la imputación de responsabilidad. Al mismo tiempo, la documentación, los registros y la transparencia exigidos por el reglamento aportan elementos probatorios que pueden paliar, aunque solo parcialmente, la opacidad técnica de los modelos avanzados de IA (Cancela-Outeda, 2024).

Como puede verse, la transparencia es el estándar que se exige en la Directiva sobre Responsabilidad en Materia de IA. Pero, tal como señala Cotino (2022), si bien la transparencia y la explicabilidad están indisolublemente unidas —las normas suelen hacer referencia a ambos conceptos a la vez—, la explicabilidad va más allá de la transparencia. La transparencia algorítmica busca dar respuesta al desequilibrio en la información entre los operadores de sistemas de inteligencia artificial y los consumidores de dicha tecnología. Por tanto, está referido al acceso a la información por parte de los usuarios cuando interactúan con este tipo de sistema.

En cambio, la explicabilidad «permite describir cómo el modelo genera predicciones. Supone que el usuario del modelo sepa cómo funciona y qué resultado se

puede esperar según las entradas. […] Los sistemas explicables pueden depurarse y supervisarse más fácilmente, y se prestan a una documentación, una auditoría y una gobernanza más completas» (Cotino, 2022: 43-44). La explicabilidad permite una mayor transparencia, pero esta no es garantía de explicabilidad, ya que en este último ámbito el usuario debe comprender los principios técnicos del sistema de IA que opera, y ello es mucho más complejo que el acceso a la información que promueve la transparencia.

De esta forma, como hemos señalado, el reenvío de la propuesta de Directiva de Responsabilidad en Materia de IA, de 2022, a las reglas de transparencia de la propuesta de Ley de Inteligencia Artificial, de 2021, es un incentivo para lograr una menor opacidad en el desarrollo de los sistemas operativos. El acceso a la información como regla para determinar si procede o no aplicar la presunción de relación de causalidad es, sin duda, un importante avance en materia de transparencia. Pero ello no necesariamente lleva a un mayor nivel de explicabilidad, por cuanto, en este último caso, lo relevante es que el la IA pueda «expresar factores importantes que influyen en los resultados del sistema de inteligencia artificial de una manera comprensible para los humanos» (Cotino, 2022: 43). De esta forma, lo importante para el nuevo régimen de responsabilidad en materia de IA es que los sistemas de IA sean comprensibles y trazables, como características de la transparencia, sin incluir la exigencia de que los algoritmos sean explicables, es decir, comparables a las justificaciones que puede dar una persona que adopta una decisión.

Conclusión

El análisis realizado permite constatar que, pese al reconocimiento normativo y ético de la explicabilidad como uno de los pilares de una IA fiable, su incidencia real en los regímenes de responsabilidad civil sigue siendo limitada. Tanto el enfoque de la Directiva 2024/2853 sobre productos defectuosos como las obligaciones técnicas del AI Act muestran que el derecho de la Unión Europea ha optado por convivir con altos niveles de opacidad técnica, compensándolos con presunciones probatorias, obligaciones de transparencia formal y acceso limitado a información relevante. En la práctica, esto significa que la falta de explicabilidad no es sancionada ni corregida estructuralmente, lo que reduce los incentivos para desarrollar sistemas de IA que sean verdaderamente comprensibles para los operadores jurídicos y para los propios usuarios.

Desde una perspectiva crítica, esta situación evidencia un desajuste entre el discurso sobre la necesidad de una IA fiable y las soluciones efectivamente adoptadas para enfrentar los problemas que la opacidad genera en el derecho de daños. La opción por mecanismos probatorios que sustituyen —pero no solucionan— la falta de explicabilidad puede resultar funcional hoy, pero es insuficiente a medida que los sis-

temas de IA avanzados ganen autonomía, complejidad y presencia en sectores de alto impacto donde la atribución clara de responsabilidades será cada vez más difícil. Por ello, resulta necesario plantear una evolución futura de la regulación. En particular, parece indispensable: i) avanzar hacia regímenes específicos de responsabilidad civil para sistemas de IA altamente autónomos; ii) reforzar los estándares de explicabilidad y trazabilidad más allá de la mera documentación técnica; iii) vincular la conformidad de los sistemas de alto riesgo a niveles verificables de interpretabilidad; y iv) promover la investigación y adopción de técnicas de la XAI que permitan reducir de forma efectiva la opacidad estructural.

De esta forma, el desafío no radica solo en adaptar el derecho a la tecnología existente, sino en orientar el desarrollo tecnológico hacia parámetros que hagan posible un modelo de responsabilidad claro y compatible con los principios básicos del derecho de daños. La opacidad técnica debe dejar de ser tolerable y la regulación deberá avanzar más allá de gestionar sus efectos para exigir condiciones que permitan comprender supervisar y justificar las decisiones automatizadas.

Referencias

AI HLEG, Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial (2019). *Ethics guidelines for trustworthy AI*. Disponible en https://tipg.link/mb2P.

Amoruso, Lucía, Mariano Bruno y Martín Dominino (2007). «Algunas diferencias entre los modelos simbolistas y conexionistas». En *XIV Jornadas de Investigación y Tercer Encuentro de Investigadores en Psicología del Mercosur* (pp. 337-338). Buenos Aires: Universidad de Buenos Aires. Disponible en https://tipg.link/mb2T.

Aneja, Urvashi (2021) «La gobernanza de la inteligencia artificial: De solucionar problemas a diagnosticarlos». *Anuario Internacional CIDOB*: 29-35. Disponible en https://tipg.link/mb2Y.

Antonov, Alexander (2022). «Gestionar la complejidad: Contribución de la UE a la gobernanza de la inteligencia artificial». *Revista CIDOB d'Afers Internacionals*, 131: 41-68. Disponible en https://tipg.link/mb2a.

Bathaee, Yavar (2018). «The artificial intelligence black box and the failure of intent and causation». *Harvard Journal of Law & Technology*, 31 (2): 890-938. Disponible en https://tipg.link/mb2q.

Cancela-Outeda, Celso (2024). «The EU's AI act: A framework for collaborative governance». *Internet of Things*, 27: 1-11. DOI: 10.1016/j.iot.2024.101291.

Cantarini, Paola (2023). «Gobernanza algorítmica, explicación por medio del diseño y justicia del diseño». *Anales de la Cátedra Francisco Suárez*, 57: 121-141. DOI: 10.30827/ACFS.v57i.25976.

Comisión Europea (2018). *Comunicación de la Comisión: Inteligencia artificial para Europa*. Disponible en https://tipg.link/mb2u.

—. (2020). *Libro blanco sobre la inteligencia artificial: Un enfoque europeo orientado a la excelencia y la confianza*. Disponible en https://tipg.link/mb2x.

—. (2021a). *Fomentar un planteamiento europeo en materia de inteligencia artificial*. Disponible en https://tipg.link/mb2-.

—. (2021b). *Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión*. Disponible en https://tipg.link/mb31.

—. (2022a). *Propuesta de Directiva relativa a la adaptación de las normas de responsabilidad civil extracontractual a la inteligencia artificial (Directiva sobre responsabilidad en materia de IA)*. Disponible en https://tipg.link/mb3A.

—. (2022b). *Nuevas normas de responsabilidad aplicables a los productos y a la IA para proteger a los consumidores y fomentar la innovación*. Disponible en https://tipg.link/mb3e.

Corrêa, Nicholas Kluge, Camila Galvão, James William Santos, Carolina del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, Rodrigo Mambrini, Luiza Galvão, Edmund Terem y Nythamar de Oliveira (2023). «Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance». *Patterns*, 4 (10): 1-14. DOI: 10.1016/j.patter.2023.100857.

Cotino, Lorenzo (2022). «Transparencia y explicabilidad de la inteligencia artificial y "compañía" (comunicación, interpretabilidad, inteligibilidad, audibilidad, estabilidad, comprobabilidad, simulabilidad.). Para qué, para quién y cuánta». En Lorenzo Cotino y Jorge Castellanos (editores), *Transparencia y explicabilidad de la inteligencia artificial* (pp. 25-69). Valencia: Tirant lo Blanch. Disponible en https://tipg.link/mb49.

Council of Europe (2024). «Framework convention on artificial intelligence and human rights, democracy and rule of law». *Council of Europe Treaty Series*, 225. Disponible en https://rm.coe.int/1680afae3c.

Da Fonseca, Ana Taveira, Elsa Vaz de Sequeira y Luís Barreto Xavier (2024). «Liability for AI driven system». En Henrique Sousa Antunes, Pedro Miguel Freitas, Arlindo L. Oliveira, Clara Martins Pereira, Elsa Vaz de Sequeira y Luís Barreto Xavier (editores). *Multidisciplinary perspectives on artificial intelligence and the law* (pp. 299-318). Springer: Cham. DOI: 10.1007/978-3-031-41264-6_16.

Dafoe, Allan (2018). *AI governance: A research agenda*. Oxford: University of Oxford.

Díaz, Regina (2007). «Responsabilidad objetiva en el ordenamiento jurídico chileno». *Revista de Derecho* (Universidad Católica del Norte), 14 (1): 79-112. DOI: 10.22199/S07189753.2007.0001.00004.

Díez-Gutiérrez, Enrique-Javier (2021). «Gobernanza híbrida digital y capitalismo EdTech: La crisis del covid-19 como amenaza». *Foro de Educación*, 19 (1): 105-133. Disponible en https://tipg.link/mb6D.

Ebers, Martin, Veronica R. S. Hoch, Frank Rosenkranz, Hannah Ruschemeier y Björn Steinrötter (2021). «The European Commission's proposal for an artificial intelligence act: A critical assessment by members of the Robotics and AI Law Society (RAILS)». *Multidisciplinary Scientific Journal*, 4 (4): 589-603. DOI: 10.3390/j4040043.

Gómez, Carlos (2022). «La propuesta de directiva sobre responsabilidad por daños causados por productos defectuosos». *InDret*, 4: 1-7. Disponible en https://tipg.link/mb6R.

Gorodnichenko, Yuriy, Tho Pham y Oleksandr Talavera (2021). «Social media, sentiment and public opinions: Evidence from #Brexit and #USElection». *European Economic Review*, 136:1-10. DOI: 10.1016/j.euroecorev.2021.103772.

Jobin, Anna, Marcello Ienca y Effy Vayena (2019). «The global landscape of AI ethics guidelines». *Nature Machine Intelligence*, 1 (9): 389-399. DOI: 10.1038/s42256-019-0088-2.

Krausová, Alžběta (2017). «Intersection between law and artificial intelligence». *International Journal of Computer*, 27 (1): 55-68. Disponible en https://tipg.link/nIh4.

Martínez, Goretty (2012). «La inteligencia artificial y su aplicación al campo del derecho». *Alegatos*, 82: 827-846. Disponible en https://tipg.link/m6xq.

Morales, Alejandro (2021). «El impacto de la inteligencia artificial en el derecho». *Advocatus*, 39: 39-71. DOI: 10.26439/advocatus2021.n39.5117.

Narváez, Camilo (2019). «La inteligencia artificial entre la culpa, la responsabilidad objetiva y la responsabilidad absoluta en los sistemas jurídicos del derecho continental y anglosajón». En Jhoel Chipana (coordinador), *Derecho y nuevas tecnologías. El impacto de una nueva era* (pp. 211-227). Lima: Themis.

Parlamento Europeo (2017). *Resolución de 17 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de derecho civil sobre robótica*. Disponible en https://tipg.link/mb8G.

—. (2020). *Resolución de 20 de octubre de 2020, con recomendaciones destinadas a la Comisión sobre un marco de los aspectos éticos de la inteligencia artificial, la robótica y las tecnologías conexas* (2020/2012 (INL)). Disponible en https://tipg.link/mb8O.

—. (2022). *Resolución de 3 de mayo de 2022, sobre la inteligencia artificial en la era digital* (2020/2266/INI). Disponible en https://tipg.link/mb8a.

Parlamento Europeo y Consejo (2024a). *Reglamento (UE) 2024/1689, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial*. Disponible en https://tipg.link/mb8_.

—. (2024b). *Directiva (UE) 2024/2853 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, sobre responsabilidad por los daños causados por productos defectuosos*. Disponible en https://tipg.link/mb96.

Parra, Darío y Ricardo Concha (2021). «Inteligencia artificial y derecho. Problemas, desafíos y oportunidades». *Vniversitas*, 70: 1-25. DOI: 10.11144/Javeriana.vj70.iadp.

—. (2022). «Responsabilidad civil por actos de robots autónomos en la Unión Europea: ¿Hacia la personalidad electrónica?». *Civilística*, 11 (3): 1-31. Disponible en https://tipg.link/mb9D.

Raposo, Vera (2024). «How is "unexplainable" and non-transparent artificial intelligence affects the rule of law: Legal and ethical challenges of black-box algorithms». *Oslo Law Review*, 11: 1-12. DOI: 10.18261/olr.11.1.6.

Ridley, Michael (2022). «Explainable artificial intelligence (XAI). Adoption and advocacy. Information technology and libraries». *Archives Journal*, 41 (2): 1-17. DOI: 10.6017/ital.v41i2.14683.

Robles, Margarita (2020). «La gobernanza de la inteligencia artificial: Contexto y parámetros generales». *Revista Electrónica de Estudios Internacionales*, 39: 1-27. Disponible en https://tipg.link/mb9K.

Russell, Stuart y Peter Norvig (2002). *Artificial intelligence: A modern approach*. 2.ª ed. Nueva Jersey: Prentice Hall.

Santana, Luis y Gonzalo Huerta (2019). «¿Son bots? Automatización en redes durante las elecciones presidenciales de Chile 2017». *Cuadernos.info*, 44: 61-77. DOI: 10.7764/cdi.44.1629.

Santos, María (2017). «Regulación legal de la robótica y la inteligencia artificial: Retos de futuro». *Revista Jurídica de la Universidad de León*, 4: 25-50. DOI: 10.18002/rjule.voi4.5285.

Solar, José (2020). «La inteligencia artificial jurídica: Nuevas herramientas y perspectivas metodológicas para el jurista». *Revus*, 41: 1-44. DOI: 10.4000/revus.6547.

Solarczyk, Alžběta (2017). «Intersections between law and artificial intelligence». *International Journal of Computer*, 27 (1): 55-68. DOI: 10.53896/ijc.v27i1.1071.

Taeihagh, Araz (2021). «Governance of artificial intelligence». *Policy and Society*, 40 (2): 137-157. DOI: 10.1080/14494035.2021.1928377.

Vale, Daniel, Ali El-Sharif y Muhammed Ali (2022). «Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law». *AI and Ethics*, 2: 815-826. DOI: 10.1007/s43681-022-00142-y.

Sobre los autores

Darío Parra Sepúlveda es abogado y profesor de Derecho Civil de la Facultad de Ciencias Jurídicas de la Universidad Austral de Chile. También es doctor en Derecho por la Universidad Carlos III de Madrid (España). Su correo electrónico es dario.parra@uach.cl 0000-0003-4024-570X.

Ricardo Concha Machuca es abogado y profesor de Derecho Civil de la Universidad de Concepción (Chile). También es doctor en Derecho por la Universidad de Chile y licenciado en Derecho por la Universidad de Concepción. Su correo electrónico es ricardo.concha@udec.cl. 0000-0002-6431-2535.

REVISTA CHILENA DE DERECHO Y TECNOLOGÍA

La *Revista Chilena de Derecho y Tecnología* es una publicación académica semestral del Centro de Estudios en Derecho, Tecnología y Sociedad de la Facultad de Derecho de la Universidad de Chile, que tiene por objeto difundir en la comunidad jurídica los elementos necesarios para analizar y comprender los alcances y efectos que el desarrollo tecnológico y cultural han producido en la sociedad, especialmente su impacto en la ciencia jurídica.

DIRECTOR
Daniel Álvarez Valenzuela
(dalvarez@derecho.uchile.cl)

SITIO WEB
rchdt.uchile.cl

CORREO ELECTRÓNICO
rchdt@derecho.uchile.cl

LICENCIA DE ESTE ARTÍCULO
Creative Commons Atribución Compartir Igual 4.0 Internacional

La edición de textos, el diseño editorial
y la conversión a formatos electrónicos de este artículo
estuvieron a cargo de Tipográfica
(www.tipografica.io).