DERECHO DIGITAL Y PROPIEDAD INTELECTUAL

7

INTELIGENCIA ARTIFICIAL Y DERECHO DE DAÑOS: CUESTIONES ACTUALES

Acorde al Reglamento (UE) 2024/1689

Itziar Alkorta Idiakez
Cristina Argelich Comelles
Maria Cristina Berenguer Albaladejo
Yolanda Bustos Moreno
Maria Raquel Evangelio Llorca
Beatriz Extremera Fernández
Pedro José Femenía López
María Remedios Guilabert Vidal
María Jorqui Azofra
Raúl Lafuente Sánchez
Pedro José López Mas
Raquel Luquin Bergareche
Andrés Marín Salmerón
Luz Martínez Velencoso
Lucía Molina Martínez
Óscar Monje Balmaseda
Esther Monterroso Casado
Juan Antonio Moreno Martínez
Carmen Muñoz García
Alberto Muñoz Villarreal
Íñigo Navarro Mendizábal
Manuel Ortiz Fernández
Miquel Peguera Poch
Antonio Rubí Puig
Alberto Tapia Hermida

Dykinson, S.L.

MORENO MARTÍNEZ, J.A.
FEMENÍA LÓPEZ, P.J.
(Coordinadores)

INTELIGENCIA ARTIFICIAL
Y DERECHO DE DAÑOS:
CUESTIONES ACTUALES

Acorde al Reglamento (UE) 2024/1689

COLECCIÓN
DERECHO DIGITAL Y PROPIEDAD INTELECTUAL

INTELIGENCIA ARTIFICIAL Y DERECHO DE DAÑOS: CUESTIONES ACTUALES

Acorde al Reglamento (UE) 2024/1689

Moreno Martínez, J.a.
Femenía López, P.j.
(Coordinadores)

Itziar Alkorta Idiakez
Cristina Argelich Comelles
Maria Cristina Berenguer Albaladejo
Yolanda Bustos Moreno
Maria Raquel Evangelio Llorca
Beatriz Extremera Fernández
Pedro José Femenía López
María Remedios Guilabert Vidal
María Jorqui Azofra
Raúl Lafuente Sánchez
Pedro José López Mas
Raquel Luquin Bergareche
Andrés Marín Salmerón
Luz Martínez Velencoso
Lucía Molina Martínez
Óscar Monje Balmaseda
Esther Monterroso Casado
Juan Antonio Moreno Martínez
Carmen Muñoz García
Alberto Muñoz Villarreal
Íñigo Navarro Mendizábal
Manuel Ortiz Fernández
Miquel Peguera Poch
Antonio Rubí Puig
Alberto Tapia Hermida

Dykinson, S.L.

Este libro ha sido sometido a evaluación por parte de nuestro Consejo Editorial.
Para mayor información, véase *www.dykinson.com/quienes_somos*

Este trabajo se enmarca en el Proyecto I+D+i (Referencia: PID2020-116185GB-100) del Ministerio de Ciencia e Innovación: "La irrupción de la inteligencia artificial en el Derecho de Daños y su adaptación a las nuevas tecnologías", siendo investigadores principales los profesores Juan Antonio Moreno Martínez y Pedro José Femenía López.

Editorial DYKINSON, S.L. Meléndez Valdés, 61 - 28015 Madrid
Teléfono (+34) 91 544 28 46 - (+34) 91 544 28 69
e-mail: info@dykinson.com
http://www.dykinson.es
http://www.dykinson.com

ISBN: 978-84-1070-708-5
Depósito Legal: M-25437-2024
DOI: https://doi.org/10.14679/3532

ISBN electrónico: 978-84-1122-801-5

Preimpresión por:
Besing Servicios Gráficos S.L.
e-mail: besingsg@gmail.com

Índice

Inteligencia artificial y responsabilidad civil: un enfoque ético en la era digital

Iñigo A. Navarro MENDIZÁBAL

Profesor Ordinario de Derecho Privado
Universidad Pontificia Comillas

1. INTRODUCCIÓN

La inteligencia artificial (IA) ha dejado de ser ciencia ficción para convertirse en una realidad omnipresente en nuestra vida diaria[1]. Desde asistentes virtuales y motores de búsqueda hasta sistemas avanzados de diagnóstico médico y vehículos autónomos, la IA está transformando la forma en que vivimos,

[1] Todo se está viendo afectado por la IA, desde el medio ambiente hasta nuestra propia mente como refleja la UNESCO: "Reconociendo las repercusiones positivas y negativas profundas y dinámicas de la inteligencia artificial (IA) en las sociedades, el medio ambiente, los ecosistemas y las vidas humanas, en particular en la mente humana, debido en parte a las nuevas formas en que su utilización influye en el pensamiento, las interacciones y la adopción de decisiones de los seres humanos», UNESCO, 2022, p.1.

trabajamos y tomamos decisiones. Este avance tecnológico trae consigo una serie de beneficios innegables, como el aumento de la eficiencia, la mejora de la precisión en diversas tareas y la capacidad de abordar problemas complejos que antes parecían insuperables.

Las ventajas de la IA las encontramos por todas partes[2]. Pero la integración masiva de la IA en la sociedad también plantea importantes desafíos éticos y legales. A medida que los sistemas de IA adquieren mayor autonomía y capacidad de toma de decisiones, surgen preguntas críticas sobre de quién es la responsabilidad y la rendición de cuentas en caso de que algo salga mal. ¿Quién es responsable si un vehículo autónomo causa un accidente? ¿Qué sucede si un algoritmo de IA discrimina injustamente a un grupo de personas? Todo esto ha generado un momento que podríamos calificar como de reflexión adversativa: sabemos las ventajas de la IA, *pero…* tememos los inconvenientes, lo que genera inseguridad y a veces miedo[3].

La ética para la IA no es solo una preocupación académica o teórica; es una necesidad práctica y urgente. La elaboración de principios éticos sólidos es crucial para garantizar que los sistemas de IA sean seguros, justos y transparentes. Los principios éticos, como el respeto a los Derechos Humanos y la

---

[2] Por ejemplo la Comisión Europea comenzaba su comunicación *Generar Confianza* diciendo: «La inteligencia artificial (IA) tiene potencial para transformar nuestro mundo para mejor: puede mejorar la asistencia sanitaria, reducir el consumo de energía, hacer que los vehículos sean más seguros y permitir a los agricultores utilizar el agua y los recursos de forma más eficiente. La IA puede utilizarse para predecir el cambio climático y medioambiental, mejorar la gestión del riesgo financiero y proporcionar las herramientas para fabricar, con menos residuos, productos a la medida de nuestras necesidades. La IA también puede ayudar a detectar el fraude y las amenazas de ciberseguridad y permite a los organismos encargados de hacer cumplir la ley luchar contra la delincuencia con más eficacia.

»La IA puede beneficiar a la sociedad y a la economía en su conjunto. Es una tecnología estratégica que se está desarrollando y utilizando a buen ritmo en todo el mundo. No obstante, también trae consigo nuevos retos para el futuro del trabajo y plantea cuestiones jurídicas y éticas», Comisión Europea, 2019, *Generar confianza*, p. 1.

[3] Por ejemplo puede verse el Informe de Ipsos Global Views on AI (2023). De acuerdo con sus resultados, la consideración de las 3 afirmaciones que siguen fueron:

— Los productos y servicios que utilizan inteligencia artificial tienen más beneficios que inconvenientes (% de acuerdo muy/algo): 54% en global y 50% en España.

— Los productos y servicios que utilizan inteligencia artificial me emocionan (% de acuerdo muy/algo): 54 % en global y 50% en España.

— Los productos y servicios que utilizan inteligencia artificial me ponen nervioso (% de acuerdo(muy/algo): 52% en global y 51% en España.

Resultados similares arrojó el estudio del Oxford Internet Institute (2020) Global Attitudes Towards AI, Machine Learning & Automated Decision Making. En sus conclusiones podemos leer: «*Internationally, sentiments about technology are ambivalent at best. There are important differences between which risks are most prominent in a particular country. For instance, North Americans and people from Western Europe see the development of AI and robotics as more likely harmful as beneficial. Survey respondents in South and East Asia are much more likely to see these developments as beneficial*».

dignidad de la persona, la seguridad y la prudencia, la solidez y la seguridad técnica, la búsqueda del bienestar y la equidad, promover la diversidad y la inclusión, la sostenibilidad y la exigencia de transparencia y responsabilidad, proporcionan un marco para guiar el desarrollo y la ejecución de la IA.

Estos principios ayudan a mitigar los riesgos asociados con la IA, como la posibilidad de errores, la perpetuación de sesgos y la falta de transparencia en la toma de decisiones. Además, promueven la confianza pública en las tecnologías de IA, lo que es esencial para su adopción generalizada y su integración exitosa en diversos ámbitos de la sociedad[4].

Por otro lado es importante considerar que los avances tecnológicos son continuos, lo que nos genera a los juristas una cierta indeterminación, porque no sabemos si seremos capaces de ofrecer una respuesta adecuada a una realidad que mañana será diferente[5]. Sin duda los tiempos de los legisladores no se parecen en nada a los tiempos de los innovadores. Como advierten Floridi et al. (2018, p. 690), la IA es una fuerza poderosa, una nueva forma de agencia inteligente, que ya está remodelando nuestras vidas, nuestras interacciones y nuestros entornos y no una mera herramienta más que habrá que regular cuando esté madura.

La ética aplicada a la IA tiene una infinidad de áreas que abordar y miles de problemas que resolver, que pueden ir desde los daños individuales y sociales causados por la tecnología y la IA hasta el debate sobre cuál es el estatus moral de los sistemas de IA o sobre cómo deberíamos tratarla en el caso de que alguna vez tuviera autoconsciencia y sentimientos. Además, el análisis ético es exhaustivo y abarca desde los comportamientos humanos más directos que diseñan, construyen o utilizan la IA, hasta el comportamiento del propio sistema de IA que puede operar con mayor o menor autonomía respecto al usuario o al diseñador o fabricante del sistema. Además, a medida que el comportamiento de un sistema de IA se vuelve más impredecible, los problemas

---

[4] Floridi et al. hablan de la ventaja de un enfoque ético «*On one side, ethics enables organisations to take advantage of the social value that AI enables*» y «*On the other side, ethics enables organisations to anticipate and avoid or at least minimise costly mistakes*», Floridi et al., 2018, p. 694.

Este enfoque dual, de forma más o menos parecida, se repite. Por ejemplo: «El objetivo de este enfoque ético es doble: por un lado, promover la concordancia o alineamiento entre las intenciones de las distintas partes y los valores éticos pertinentes para el uso previsto; por otro lado, identificar, corregir o denunciar las aplicaciones que sirven a fines éticamente inaceptables que ignoran, o violan, valores decisivos en relación con su ámbito de actuación», Llano Alonso, 2024, p. 188.

[5] Como Susskind escribe muy gráficamente, no hay línea de llegada: "*what at once excites and unsettles me most is that there is no finishing line. No-one in Silicon Valley, China, or South Korea is dusting their hands off, proclaiming, 'job done'*", Susskind, 2019, p. 41.

éticos aumentan y la rendición de cuentas debe replantearse[6], siendo también diferentes los problemas que plantean la IA específica y la IA general[7].

En este contexto, el objetivo de este capítulo es explorar en profundidad la intersección entre la responsabilidad civil (RC) y los principios éticos de la IA cuando ésta causa daños. A través de un análisis detallado, pretende proporcionar una visión clara y comprensiva de cómo los principios éticos pueden y deben influir en la asignación de responsabilidad legal en casos de daños causados por la IA. También busca contribuir al debate académico y práctico sobre cómo manejar de manera efectiva los desafíos y oportunidades que presenta la IA. La combinación de un enfoque ético con una sólida base legal es esencial para asegurar que el avance tecnológico sea seguro, justo y equitativo para todos.

2. PRINCIPIOS ÉTICOS DE LA IA

2.1. La importancia de la Ética en la IA

Como señalaba la Comisión Europea (2019, p. 10) en su comunicación *Generar confianza*: «La dimensión ética de la IA no es un lujo ni un algo accesorio: ha de ser parte integrante del desarrollo de la IA. Al tratar de lograr una IA centrada en el ser humano basada en la confianza, salvaguardamos el respeto de los valores esenciales de nuestra sociedad y forjamos una marca distintiva para Europa y su industria como líder de la IA de vanguardia en la que se puede confiar en todo el mundo».

En esta misma comunicación, la Comisión Europea declara «La UE se asienta sobre un sólido marco normativo, que constituirá la referencia mundial para la IA centrada en el ser humano» convirtiendo a la UE en el principal regulador que además está sirviendo de inspiración para muchos otros países (efecto Bruselas) [8].

---

[6] Dignum, 2018, p. 1.

[7] Se podría hablar de distintas aproximaciones al análisis ético de la IA, todos ellos necesarios (Dignum, 2018, p. 2):

- "*Ethics by Design: the technical/algorithmic integration of ethical reasoning capabilities as part of the behaviour of artificial autonomous system;*

- *Ethics in Design: the regulatory and engineering methods that support the analysis and evaluation of the ethical implications of AI systems as these integrate or replace traditional social structures;*

- *Ethics for Design: the codes of conduct, standards and certification processes that ensure the integrity of developers and users as they research, design, construct, employ and manage artificial intelligent systems*".

[8] El concepto «efecto Bruselas» lo expuso Bradford analizando el impacto global del poder normativo de la UE y cómo sus regulaciones se extienden a nivel mundial sin necesidad de instituciones internacionales ni acuerdos de cooperación explícitos (Bradford, 2012).

La necesidad de una regulación basada en principios éticos es urgente en la UE, porque si no dictamos las reglas, nos serán dictadas[9]. Además es necesario que la regulación tenga el «enfoque europeo» basado en principios éticos[10].

Dentro de esta importancia de la ética, Mittelstadt et al. (2016) en su interesante artículo *The ethics of algorithms: Mapping the debate*, elaboraron un mapa conceptual de las principales preocupaciones éticas que nos presentan los algoritmos que toman decisiones y que afectan a la sociedad y a los individuos. En el artículo se identifican:

A) Tres tipos de preocupaciones epistémicas:

1. La evidencia inconclusa: el conocimiento probable y a menudo incierto que producen los algoritmos puede llevar a acciones basadas en información insuficiente.
2. La evidencia inescrutable: la falta de transparencia en cómo los algoritmos generan conclusiones puede hacer que las decisiones sean difíciles de entender y escrutar.
3. La evidencia errónea de la IA cuando la calidad de las conclusiones de los algoritmos está limitada por la calidad de los datos de entrada.

B) Dos tipos de preocupaciones normativas:

1. Los resultados injustos cuando las acciones impulsadas por algoritmos pueden ser evaluadas según su justicia, afectando a ciertos grupos de manera desproporcionada.
2. Los efectos transformadores, porque los algoritmos pueden cambiar cómo conceptualizamos el mundo y reorganizar aspectos sociales y políticos.

C) La trazabilidad, que es una preocupación transversal y que está relacionada con la capacidad de asignar responsabilidad y rendir cuentas por los efectos de los algoritmos.

---

[9] Reconoce el Parlamento Europeo: «si la Unión no actúa con rapidez y valentía, acabará teniendo que seguir las reglas y normas fijadas por otros y corre el riesgo de sufrir efectos perjudiciales para la estabilidad política, la seguridad social, los derechos fundamentales, las libertades individuales y la competitividad económica », Parlamento Europeo, 2020, *IA en la Era Digital*, 6.

[10] «Pone de relieve que un objetivo subyacente a la estrategia digital de la Unión, así como al de la estrategia de la IA, es crear un «enfoque europeo» en un mundo digitalizado; aclara que este enfoque debe estar centrado en el ser humano, ser fiable, guiarse por principios éticos y basarse en el concepto de economía social de mercado; subraya que la persona y la protección de sus derechos fundamentales deben permanecer siempre en el centro de todas las consideraciones políticas», Parlamento Europeo, 2020, *IA en la Era Digital*, 130.

Quizás una de las cuestiones que muestra de manera clara la necesidad de la ética de la IA, es que los algoritmos toman decisiones por las personas. Así, pueden decidir tanto desde qué trayecto seguir para llegar al destino, qué restaurante visitar o con qué persona quedar para cenar, en el ámbito personal, como sobre cuestiones empresariales (inversiones, contrataciones entre muchas otras) gobernando cada vez más las empresas y las instituciones. Como señalan Mittelstadt et al. (2016, p. 1), «las operaciones, las decisiones y las opciones que antes se dejaban en manos de los seres humanos se delegan cada vez más en los algoritmos, que pueden aconsejar, si no decidir, sobre cómo deben interpretarse los datos y qué acciones deben adoptarse como resultado».

Pero además de esta "delegación de funciones", cada vez es más difícil saber cómo toman las decisiones los algoritmos cuya incertidumbre y opacidad es cada vez más problemática[11]. Ante esta capacidad de aprender de los algoritmos hay incluso quienes piensan que deben ser considerados agentes morales con una cierta responsabilidad moral[12].

La autonomía de la IA es una de las razones que impone con más fuerza la necesidad de una reflexión ética[13], porque al poder adoptar decisiones sin un control humano directo, se puede perder el control ético generalmente atribuido a las personas. Esto está en el fondo de muchas de las distopías que se han bosquejado en la ciencia ficción y que podrían resumirse en la consideración de Bostrom y Yudkowsky (2014) según la cual una IA superinteligente, a menos que estuviera diseñada específicamente para actuar de manera beneficiosa para los humanos, podría perseguir sus propios objetivos de manera perjudicial para la humanidad, incluida la extinción humana[14].

Muchos de estos temores gravitan en torno a las ideas de:

— La desalienación de objetivos entre los humanos y la IA[15]: «*The complexity of human values and the difficulty of embedding them correctly*

[11] Mittelstadt et al., 2016, p. 3.

[12] «*For some, learning algorithms should be considered moral agents with some degree of moral responsibility*», Mittelstadt et al., 2016, p.11.

Este debate no lo he tratado en este capítulo por múltiples razones, siendo una de las principales que en la UE se ha ido "desinflando" la idea de una posible RC de la IA o de los robots que tuvieran "personalidad digital" y se ha optado por la RC de personas tales como el fabricante.

[13] «*Incorporating ethical principles into the design of AI systems is essential to ensure that their decision-making processes are transparent and that they can be audited and controlled effectively*», Bostrom y Yudkowsky, 2014.

[14] «*A superintelligent AI, unless specifically designed to act in ways that are beneficial to humans, could pursue its own goals in ways that are harmful to humanity, including causing human extinction*», Bostrom y Yudkowsky, 2014.

[15] Por ejemplo, estos miedos están relacionados con los principios 10 y 11 de los Principios Asilomar: «10) *Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation*» y «11) *Human Values: AI*

in a machine makes it likely that any simple goal given to an AI could result in unforeseen and potentially disastrous actions» (Bostrom y Yudkowsky, 2014) [16].

— El miedo a la pérdida de control: «*Ensuring that an AI remains under human control is essential, as an uncontrolled AI could modify itself in ways that make it even more difficult to control, potentially leading to catastrophic outcomes*» (Bostrom y Yudkowsky, 2014).

La integración de los principios éticos en normativas y estándares es crucial para garantizar que las tecnologías de IA se desarrollen y utilicen de manera ética[17]. Quizás una argumentación poderosa que demuestra la necesidad de la ética en este campo sería la *reductio ad absurdum*, pues ¿qué pasaría si se desarrollara la IA sin ninguna cortapisa? Probablemente nos acercaríamos en mayor o menor grado a las distopías de la ciencia ficción. Podrían ocurrir cosas tales como:

— Amenazas a la seguridad tanto individual como colectiva. La IA podría usarse para fines maliciosos como el desarrollo sin control de armas autónomas o la manipulación masiva de la información para desestabilizar sociedades. Es más, la IA podría autónomamente actuar de manera impredecible y peligrosa si no respetara principios éticos, lo que podría llevar a graves amenazas para la seguridad global.

— Deshumanización. La IA que toma decisiones, si no tuviera un control ético se deshumanizaría, lo que podría suponer no respetar valores tales como la dignidad o los derechos fundamentales y llevar una pérdida de control que incluso podría poner en riesgo la existencia de la humanidad.

— Automatización de la injusticia, pues se podrían perpetuar o programar sesgos con el fin de que se tomaran decisiones discriminatorias en áreas tan sensibles como la justicia, el mercado laboral o bancario o la asignación de recursos.

---

systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity», Future of Life Institute, 2017.

[16] Véase también Llano Alonso, 2024, pp. 184-187.

[17] También existen otras opiniones alternativas como la de Müller que analizando el problema de la alineación de valores en la IA propone, más allá de intentar crear «máquinas morales», el enfoque de la «domesticación» con una analogía con la relación que los seres humanos tenemos con los animales no humanos: «*while nonhuman animals and AI agents are different in many ways, the fundamental approach with which we have "value-aligned" animals is transferable to AI agents, and can guide us in understanding how human moral agents can cooperate with nonhuman agents without human-like moral capabilities*», Müller, 2022, p. 235.

— Neoludismo. Una IA sin ética también podría generar una masiva reacción en contra de la tecnología como un neoludismo debido a que se generaría una falta de confianza púbica.

Muchos organismos internacionales han comenzado a adoptar marcos regulatorios y directrices para orientar el uso ético de la IA. Se pueden citar, como principales:

— **Directrices Éticas para una IA fiable**[18]**:** establece un marco normativo y ético para el desarrollo y uso de la IA en la UE. Estas directrices buscan garantizar que la IA sea confiable, segura y respetuosa con los derechos fundamentales y los valores europeos. Su importancia radica en promover un enfoque de la IA que no solo impulse la innovación, sino que también proteja a los individuos y la sociedad, fomentando la transparencia, la responsabilidad y la sostenibilidad en la adopción de estas tecnologías[19].

---

[18] Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial. (2019). *Directrices éticas para una IA fiable*. Comisión Europea.

Las Directrices Éticas para una IA Fiable fueron elaboradas por el Grupo Independiente de Expertos de Alto Nivel sobre IA, un comité creado por la Comisión Europea en 2018 (en la bibliografía y en las citas he considerado autor al Grupo, aunque el autor corporativo es la propia Comisión Europea). El proceso de elaboración de las directrices siguió un enfoque abierto, inclusivo y deliberativo y se pueden citar como momentos clave:

— creación del grupo de expertos: la Comisión Europea seleccionó a un grupo multidisciplinario de expertos procedentes de la academia, la industria y la sociedad civil. Estos expertos tenían conocimientos diversos en áreas como IA, derecho, filosofía, ética, y economía, asegurando una amplia representación de diferentes perspectivas;

— proceso de consulta pública que fue uno de los pilares de la elaboración: tras elaborar un borrador se abrió una consulta pública en la que ciudadanos, investigadores, ONGs, y empresas pudieron compartir sus opiniones y preocupaciones respecto a la IA y su desarrollo ético;

— revisión en varias fases con ajuste del documento tras los comentarios recibidos;

— publicación de la versión definitiva el 8 de abril de 2019.

[19] Como se verá a lo largo del texto, considero que estas Directrices son una base fundamental dentro de la UE para hacer un análisis ético de la IA que sirva para fundamentar la regulación.

Esto no obstante, ha habido diversas posiciones críticas con este documento que en aras a la objetividad conviene citar. Por ejemplo Stamboliev y Christiaens consideran que el concepto de *Trustworthy AI* (IA Confiable) funciona como un «significante vacío», es decir, un término vago que busca unir a diferentes actores con intereses contradictorios, como la industria, los expertos en ética y los legisladores, si bien lo hacen tomando conceptos de la filosofía política de los filósofos postmarxistas Ernesto Laclau y Chantal Mouffe. Los autores del artículo destacan la subordinación de la ética a la industria tecnológica adoptando aquella el rol de un "extintor de incendios" para minimizar daños una vez que la tecnología está desarrollada. Entre los posibles escenarios futuros que plantean están: (i) *Status Quo*, en el que se mantiene el equilibrio frágil actual; (ii) *Asimilación*, donde la ética queda completamente absorbida por los intereses de la industria; y (iii) *Contestación*, en el que la ética desafía la hegemonía de la industria y toma un rol más crítico y autónomo (Stamboliev y Christiaens, 2024). Considero que quizás podríamos pensar en una asimilación con contestación, en la que la ética manteniendo una cierta independencia, se integrara en la regulación que la industria debe cumplir.

Este documento no es vinculante, pero, dentro del llamado «enfoque europeo»[20], proporciona un marco de referencia ético para el desarrollo de la IA dentro de la UE, basado en los valores fundamentales de la UE, tales como el respeto a los derechos humanos, la democracia, el Estado de derecho y el bienestar social. Pretende ser el pilar ético de la normativa, por lo que lógicamente influirá en futuras legislaciones y políticas sobre IA, por lo que es particularmente importante para nosotros.

— **Recomendación sobre la ética de la inteligencia artificial de la UNESCO**[21]**:** adoptada en 2021establece principios y valores fundamentales para guiar el desarrollo y uso de la IA en todo el mundo: subraya la importancia de la dignidad humana, la sostenibilidad, la diversidad y la inclusión en la IA. Propone un marco ético que abarca desde la transparencia y la rendición de cuentas hasta la protección de datos y la equidad. Su objetivo es garantizar que la IA se desarrolle y aplique de manera que respete los derechos humanos y promueva el bienestar de todos, evitando cualquier tipo de discriminación o daño social. Esta recomendación es relevante porque ofrece una guía globalmente reconocida para la adopción responsable de la IA, influenciando políticas y prácticas a nivel internacional.

— **Principios de la OCDE sobre IA**[22]**:** buscan promover un desarrollo y uso de la IA que sea innovador y responsable. Estos principios se centran en asegurar que la IA respete los derechos humanos, promueva el crecimiento inclusivo, sea transparente y explicable, y esté diseñada para funcionar de manera robusta y segura. Además, subrayan la importancia de la cooperación internacional y la necesidad de establecer marcos regulatorios que promuevan la confianza en la IA a nivel global. Su relevancia radica en la orientación que brindan a los gobiernos y las empresas para que el desarrollo de la IA se realice de manera ética y beneficie a la sociedad en su conjunto.

— **Principios de Asilomar para la IA**[23]**:** son un conjunto de principios formulados en 2017 por expertos en IA, ética y tecnología durante una conferencia en Asilomar, California. Estos principios están diseñados

---

También ha habido críticas que acusan al documento de promover el *ethic washing*, como recoge Larsson (2020, p. 443).

[20] Parlamento Europeo, 2020, *IA en la Era Digital*, 130.

[21] UNESCO. (2022). *Recomendación sobre la ética de la inteligencia artificial*. La Recomendación sobre la ética de la inteligencia artificial de la UNESCO fue adoptada por los 193 Estados Miembros de la Organización en noviembre de 2021.

[22] OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449.

[23] Future of Life Institute. (2017). *Asilomar AI Principles*.

para garantizar que el desarrollo de la inteligencia artificial sea seguro, ético y beneficioso para la humanidad. Los principios incluyen pautas sobre la transparencia, la justicia, la responsabilidad y la alineación con los valores humanos, así como la necesidad de una investigación robusta y una cooperación internacional para mitigar los riesgos asociados con la IA avanzada. La importancia de los Principios de Asilomar radica en su enfoque proactivo para abordar los desafíos éticos y de seguridad que plantea la IA, promoviendo un desarrollo de la tecnología que sea responsable y alineado con los intereses del bienestar global.

— **Declaración de Montreal**[24]: creada en 2018 por un grupo de académicos, expertos y miembros de la sociedad civil en Montreal, esta Declaración busca asegurar que la IA respete la dignidad humana, los derechos individuales y el bienestar colectivo. Los principios clave incluyen la transparencia, la equidad, la responsabilidad, la inclusión y la sostenibilidad. La carta también subraya la necesidad de que las decisiones automatizadas sean comprensibles y explicables, y que las tecnologías de IA se utilicen para el beneficio común, evitando sesgos y discriminaciones. Su importancia radica en ofrecer un marco ético que guíe a desarrolladores, empresas y gobiernos en la adopción de la IA de manera que promueva el respeto por los valores humanos y la justicia social.

— **Directrices del IEEE para un Diseño Ético de la IA**[25]**:** recomendaciones desarrolladas por el Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) para orientar el diseño y la implementación de la IA y los sistemas autónomos. Estas directrices enfatizan la necesidad de integrar principios éticos desde las primeras etapas del desarrollo de la IA, promoviendo la protección de los derechos humanos, la privacidad, la equidad y la transparencia. También abordan cuestiones como la responsabilidad, la seguridad, y el impacto social de la tecnología. La importancia de estas directrices radica en su enfoque en crear sistemas de IA que no solo sean técnicamente robustos, sino también alineados con los valores éticos y las expectativas sociales, ayudando a minimizar los riesgos y maximizar los beneficios de la IA para todos los usuarios.

---

[24] Institut de Valorisation Des Données (IVADO). (2018). *Montreal Declaration for a Responsible Development of Artificial Intelligence.*

[25] IEEE. (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition.*

También hay muchísimas otras iniciativas como la ***Guía de buenas prácticas en el uso de la IA ética de OdiseIA***[26] que proporciona directrices y recomendaciones para que las organizaciones y empresas implementen sistemas de IA de manera ética y responsable, siguiendo principios éticos y normativos[27].

2.2. Principales principios éticos

Como se ha visto son muchos los documentos que han tratado de buscar unos principios éticos para la IA partiendo de diferentes enfoques y/o analizando distintas áreas.

En este epígrafe se han analizado dichos documentos con el objetivo de extraer los principios éticos que deben gobernar la IA para aplicarlos posteriormente a la RC.

A continuación se presenta los principales:

2.2.1. Derechos Humanos y Dignidad

Que el respeto a los Derechos humanos y a la dignidad de la persona es un principio que debe ordenar el desarrollo y uso de la IA es una obviedad elemental que no habría ni que mencionar. Sin embargo, en muchos de los documentos que hemos aludido, aparece citado explícitamente como uno de los principios, por lo que conviene señalarlo (lo que abunda no daña reza el refranero español).

Más que un principio ético podría decirse que es el cimiento sobre el que se construye el resto de la arquitectura ética. Como señalan las Directrices

---

[26] OdiseIA, PwC, Google, Microsoft, IBM, y Telefónica. (2022). *Guía de buenas prácticas en el uso de la inteligencia artificial ética.*

[27] No faltan en la literatura opiniones que se muestran pesimistas en cuanto a la efectividad práctica de las guías éticas que se están desarrollando sobre la IA. Por ejemplo Hagendorff, en su trabajo *The Ethics of AI Ethics: An Evaluation of Guidelines*, realiza un análisis de 22 guías éticas sobre IA, poniendo en evidencia su limitada efectividad para influir en la toma de decisiones de los desarrolladores y la industria. A pesar de que estas guías formulan principios como la transparencia, la privacidad y la justicia, Hagendorff destaca la falta de mecanismos de aplicación reales que aseguren su cumplimiento. Este vacío, sostiene, permite a las empresas utilizar las guías como herramientas de relaciones públicas, desalentando la creación de un marco legal vinculante. Además, las guías tienden a centrarse en soluciones técnicas, mientras omiten cuestiones más amplias como los costos sociales y ecológicos, así como los riesgos de abuso político de la IA. Para mitigar estas deficiencias, el autor aboga por complementar las guías deontológicas con un enfoque basado en la ética de las virtudes, que fomente la responsabilidad moral en los desarrolladores. Estas reflexiones nos deben hacer pensar para fundamentar un marco de RC que exija rendición de cuentas ante posibles daños derivados del uso indebido o negligente de la IA, Hagendorff, 2020, pp. 99-120.

éticas para una IA fiable «El respeto de los derechos fundamentales, dentro de un marco de democracia y estado de Derecho, proporciona la base más prometedora para identificar los principios y valores éticos abstractos que se pueden poner en práctica en el contexto de la IA» [28].

Por otro lado también nos recuerda con claridad que la legalidad, el Derecho y la Ética no difieren cuando estamos en la realidad *offline* y *online*[29]. En ocasiones parece que algunas personas no consideran de la misma manera lo que ocurre en el mundo digital y lo que pasa fuera de él, pero un daño moral o una filtración de datos que genere daños resarcibles se puede producir en un periódico tradicional, en Instagram o en X.

Dentro de este apartado se pueden señalar tres ejes:

A) Respeto de los Derechos humanos

Los sistemas de IA deben respetar y promover los derechos fundamentales, garantizando que su desarrollo y uso no vulnere la dignidad, las libertades y los derechos de las personas. Ejemplo de su formulación puede ser:

> «La dignidad inviolable e intrínseca de cada ser humano constituye la base del sistema universal, indivisible, inalienable, interdependiente e interrelacionado de derechos humanos y libertades fundamentales. Por consiguiente, el respeto, la protección y la promoción de la dignidad humana y de los derechos establecidos por el derecho internacional, en particular el derecho internacional de los derechos humanos, son esenciales a lo largo del ciclo de vida de los sistemas de IA» (UNESCO, 2021, 13) [30].

---

[28] Grupo Independiente de Expertos, 2019, *Directrices éticas*, p. 12.

[29] Por ejemplo: «Los Estados Miembros deberían velar por que se investiguen y reparen los daños causados mediante sistemas de IA, estableciendo mecanismos de aplicación estrictos y medidas correctivas, a fin de asegurarse de que los derechos humanos, las libertades fundamentales y el estado de derecho son respetados en el mundo digital y en el mundo físico» (UNESCO, 2022, p. 55).

[30] Otras formulaciones:

- «*A/IS should be designed and operated in a way that both respects and fulfills human rights, freedoms, human dignity, and cultural diversity*». (IEEE, 2019, Principle 1, p. 22).

- «*AI actors should respect the rule of law, human rights, democratic and human-centred values throughout the AI system lifecycle. These include non-discrimination and equality, freedom, dignity, autonomy of individuals, privacy and data protection, diversity, fairness, social justice, and internationally recognised labour rights. This also includes addressing misinformation and disinformation amplified by AI, while respecting freedom of expression and other rights and freedoms protected by applicable international law*», OECD, 2019, p. 8.

— Principios Asilomar: «11) *Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity*», Future of Life Institute, 2017.

B) Respeto de la Autonomía humana

La IA debe permitir que las personas mantengan el control sobre sus decisiones y acciones en todos los ámbitos:

— «Las personas que interactúen con sistemas de IA deben poder mantener una autonomía plena y efectiva sobre sí mismas y ser capaces de participar en el proceso democrático. Los sistemas de IA no deberían subordinar, coaccionar, engañar, manipular, condicionar o dirigir a los seres humanos de manera injustificada. En lugar de ello, los sistemas de IA deberían diseñarse de forma que aumenten, complementen y potencien las aptitudes cognitivas, sociales y culturales de las personas. La distribución de funciones entre los seres humanos y los sistemas de IA debería seguir principios de diseño centrados en las personas, y dejar amplias oportunidades para la elección humana», Grupo Independiente de Expertos, 2019, *Directrices éticas*, 50, p. 15.

— «Los sistemas de IA deberían respaldar la autonomía y la toma de decisiones de las personas, tal como prescribe el principio del *respeto de la autonomía humana*. Esto requiere que los sistemas de IA actúen tanto como facilitadores de una sociedad democrática, próspera y equitativa, apoyando la acción humana y promoviendo los derechos fundamentales, además de permitir la supervisión humana», Grupo Independiente de Expertos, 2019, *Directrices éticas* 62, p. 19[31].

C) Respeto de la privacidad e intimidad

La protección de la privacidad y la intimidad es esencial en el contexto de la IA. Este principio establece que los sistemas no deben invadir la vida privada ni recopilar datos sin un control estricto.

La privacidad de los datos personales debe ser protegida en todas las etapas del desarrollo y uso de sistemas de IA. Esto incluye la recopilación, almacenamiento y procesamiento de datos. Los desarrolladores deben implementar medidas de seguridad robustas y asegurar que los datos se utilicen de manera

---

[31] Otras enunciaciones:

— Declaración de Montreal: «*AIS must allow individuals to fulfill their own moral objectives and their conception of a life worth living*» (IVADO, 2018, Principio 2: Respeto a la Autonomía, p. 9).

— Principios Asilomar «13) *Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty*».

— «16) *Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives*», Future of Life Institute, 2017.

ética y con el consentimiento informado de los individuos. Por ejemplo, los asistentes virtuales deben garantizar que las conversaciones privadas no se almacenen ni se compartan sin el consentimiento del usuario.

Ejemplo:

> «La privacidad es un derecho fundamental que se ve especialmente afectado por los sistemas de IA, y que guarda una estrecha relación con el *principio de prevención del daño*. La prevención del daño a la privacidad también requiere una adecuada gestión de los datos, que abarque la calidad y la integridad de los datos utilizados, su pertinencia en contraste con el ámbito en el que se desplegarán los sistemas de IA, sus protocolos de acceso y la capacidad para procesar datos sin vulnerar la privacidad». (Directrices éticas para una IA fiable, 71, p. 21) [32].

2.2.2. Seguridad y Prudencia. Solidez y seguridad técnica

Los sistemas de IA deben ser seguros y robustos, garantizando su correcto funcionamiento durante toda su vida útil, y deben prever mecanismos para minimizar riesgos. Los desarrolladores de IA deben anticipar posibles consecuencias negativas y tomar medidas para evitarlas[33], promoviendo un desarrollo seguro y controlado. Además, los sistemas de IA deben ser diseñados para minimizar los riesgos de mal uso o abuso, implementando salvaguardas apropiadas.

---

[32] Otras exposiciones:

- Declaración de Montreal: «*Personal spaces in which people are not subjected to surveillance or digital evaluation must be protected from the intrusion of AIS and data acquisition and archiving systems (DAAS). The intimacy of thoughts and emotions must be strictly protected from AIS and DAAS uses capable of causing harm, especially uses that impose moral judgments on people or their lifestyle choices*» (IVADO, 2018, Principio 3: Protección de la Privacidad e Intimidad, p. 10).

- Principios Asilomar «12) *Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data*», Future of Life Institute, 2017.

[33] Por ejemplo: «*Plan de repliegue y seguridad general.* Los sistemas de IA deberían contar con salvaguardias que posibiliten un plan de repliegue en el caso de que surjan problemas. Esto puede significar que los sistemas de IA pasen de un procedimiento basado en estadísticas a otro basado en normas, o que soliciten la intervención de un operador humano antes de proseguir con sus actuaciones.39 Es preciso garantizar que el sistema se comportará de acuerdo con lo que se espera de él sin causar daños a los seres vivos ni al medio ambiente. Esto incluye la minimización de las consecuencias y errores imprevistos. Además, se deberían establecer procesos dirigidos a aclarar y evaluar los posibles riesgos asociados con el uso de sistemas de IA en los diversos ámbitos de aplicación. El nivel de las medidas de seguridad requeridas depende de la magnitud del riesgo que plantee un sistema de IA, que a su vez depende de las capacidades del sistema. Cuando se prevea que el proceso de desarrollo o el propio sistema planteará riesgos particularmente altos, es crucial desarrollar y probar medidas de seguridad de forma proactiva», Grupo Independiente de Expertos, 2019, *Directrices éticas*, 68.

> «2. Solidez técnica y seguridad. 66) Un componente crucial de la IA fiable es la solidez técnica, que está estrechamente vinculada al *principio de prevención del daño*. La solidez técnica requiere que los sistemas de IA se desarrollen con un enfoque preventivo en relación con los riesgos, de modo que se comporten siempre según lo esperado y minimicen los daños involuntarios e imprevistos, evitando asimismo causar daños inaceptables. Lo anterior debería aplicarse también a los cambios potenciales en su entorno operativo o a la presencia de otros agentes (humanos y artificiales) que puedan interactuar con el sistema de manera contenciosa. Además, debería garantizarse la integridad física y mental de los seres humanos», (Directrices éticas para una IA fiable, 66, p. 20) [34].

2.2.3. Bienestar y Equidad. Diversidad e Inclusión

Los sistemas de IA deben contribuir al bienestar humano[35], mejorando la calidad de vida, la salud y las condiciones laborales, sin causar daño. **Priorizar el bienestar humano** debe ser un objetivo principal del diseño de la

---

[34] Otras expresiones:

- UNESCO: «Seguridad y protección 27. Los daños no deseados (riesgos de seguridad) y las vulnerabilidades a los ataques (riesgos de protección) deberían ser evitados y deberían tenerse en cuenta, prevenirse y eliminarse a lo largo del ciclo de vida de los sistemas de IA para garantizar la seguridad y la protección de los seres humanos, del medio ambiente y de los ecosistemas. La seguridad y la protección de la IA se propiciarán mediante el desarrollo de marcos de acceso a los datos que sean sostenibles, respeten la privacidad y fomenten un mejor entrenamiento y validación de los modelos de IA que utilicen datos de calidad.», UNESCO, 2022, p. 10.

- Principios de Asilomar «*AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible*», (Future of Life Institute, 2017, Principio 6: Safety).

- Declaración de Montreal: «*Every person involved in AI development must exercise caution by anticipating, as far as possible, the adverse consequences of AIS use and by taking the appropriate measures to avoid them*», (IVADO, 2018, Principio 8: Prudencia, p. 15).

- OECD: «1.4. *Robustness, security and safety a) AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety and/or security risks.*

b) Mechanisms should be in place, as appropriate, to ensure that if AI systems risk causing undue harm or exhibit undesired behaviour, they can be overridden, repaired, and/or decommissioned safely as needed.

c) Mechanisms should also, where technically feasible, be in place to bolster information integrity while ensuring respect for freedom of expression», OECD, 2019, p. 9.

- IEEE: «*Mitigate risks and negative impacts, including misuse, as A/IS evolve as socio-technical systems. In particular by ensuring A/IS are accountable and transparent*», IEEE, 2019, p. 20.

[35] «*The principle of creating AI technology that is beneficial to humanity is expressed in different ways, but it typically features at the top of each list of principles. Montreal and IEEE principles both use the term 'well-being': for Montreal, 'the development of AI should ultimately promote the well-being of all sentient creatures'; while IEEE states the need to 'prioritize human well-being as an outcome in all system designs*», Floridi et al., 2018, p. 696.

IA, utilizando métricas ampliamente aceptadas para medir el impacto en la calidad de vida.

> «Para ello, es necesario que los sistemas de IA se centren en las personas y se fundamenten en el compromiso de utilizarlos al servicio de la humanidad y del bien común, con el objetivo de mejorar el bienestar y la libertad de los seres humanos» (Grupo Independiente de Expertos, 2019, *Directrices éticas*, 10, p. 5) [36].

La IA debe **promover la Equidad** y garantizar que sus beneficios se distribuyan de manera justa, reduciendo las desigualdades sociales y económicas, sin perpetuar sesgos o discriminaciones. El principio de equidad también implica que el desarrollo de la IA debe promover un crecimiento inclusivo y sostenible, asegurando la igualdad de oportunidades y acceso.

La justicia en la IA implica la equidad en la distribución de beneficios y riesgos, evitando cualquier forma de discriminación o sesgo. Los sistemas de IA deben ser diseñados y entrenados con datos diversos y representativos para asegurar que no favorezcan a ningún grupo en detrimento de otro. Por ejemplo, los algoritmos de contratación deben ser evaluados para garantizar que no discriminen a candidatos por motivos de género, raza o edad y que no se perpetúen o amplíen las desigualdades existentes. La justicia puede influir en la responsabilidad civil al requerir que los sistemas de IA no discriminen ni causen daños desproporcionados a grupos vulnerables.

> «El desarrollo, despliegue y utilización de sistemas de IA debe ser equitativo. Pese a que reconocemos que existen muchas interpretaciones diferentes de la equidad, creemos que esta tiene tanto una dimensión sustantiva como procedimental. La dimensión sustantiva implica un compromiso de: garantizar una distribución justa e igualitaria de los beneficios y costes, y asegurar que las personas y grupos no sufran sesgos injustos, discriminación ni estigmatización. Si se

---

[36] Otras muestras:

Declaración de Montreal «*AIS must help individuals improve their living conditions, their health, and their working conditions*», IVADO, 2018, Principio 1: Bienestar, p. 8.

- Principios Asilomar «1) *Research Goal: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence*», Future of Life Institute, 2017.

- IEEE: «*A/IS should prioritize human well-being as an outcome in all system designs, using the best available and widely accepted well-being metrics as their reference point*» IEEE, 2019, p. 241.

- OECD: «1.1. *Inclusive growth, sustainable development and well-being. Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, well-being, sustainable development and environmental sustainability*», OECD, 2019, p. 8.

> pueden evitar los sesgos injustos, los sistemas de IA podrían incluso aumentar la equidad social. También se debería fomentar la igualdad de oportunidades en términos de acceso a la educación, los bienes los servicios y la tecnología. Además, el uso de sistemas de IA no debería conducir jamás a que se engañe a los usuarios (finales) ni se limite su libertad de elección. Asimismo, la equidad implica que los profesionales de la IA deberían respetar el principio de proporcionalidad entre medios y fines, y estudiar cuidadosamente cómo alcanzar un equilibrio entre los diferentes intereses y objetivos contrapuestos. La dimensión procedimental de la equidad conlleva la capacidad de oponerse a las decisiones adoptadas por los sistemas de IA y por las personas que los manejan, así como de tratar de obtener compensaciones adecuadas frente a ellas. Con este fin, se debe poder identificar a la entidad responsable de la decisión y explicar los procesos de adopción de decisiones», (Grupo Independiente de Expertos, 2019, *Directrices éticas*, 52, pp. 15 y 16) [37].

La IA debe respetar y promover la **diversidad** cultural y social, evitando la homogeneización de comportamientos y opiniones, y asegurando que las diferencias sean respetadas[38]. Además, el desarrollo de la IA debe ser **inclusivo**,

---

[37] Otros ejemplos:

- Declaración de Montreal: «*AIS development must produce social and economic benefits for all by reducing social inequalities and vulnerabilities*», IVADO, 2018, Principio 6: Equidad, p. 13.

- OECD: «*Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, well-being, sustainable development and environmental sustainability*», OCDE, 2019, p. 8.

[38] Hay que tener en cuenta que algunas de las tendencias de la IA deben ser analizadas con detenimiento. Por ejemplo, Colmenarejo et al. (2022, pp. 114-115) reflejan la tensión que existe, entre otras, entre la estandarización y la localización. A través de la armonización de estándares, los sistemas de IA podrían ser conformes con la normativa, lo que ayudaría a la certeza jurídica y sería bueno para el mercado único. Sin embargo, una estandarización demasiado rígida podría no captar adecuadamente las complejidades y peculiaridades locales de los distintos contextos socioculturales. Esto podría generar soluciones de "talla única" que no abordaran problemas específicos, como ciertas desigualdades o sensibilidades culturales que varían entre regiones.

Por el contrario, la localización implica adaptar los sistemas y las regulaciones de IA a los contextos específicos de cada área o comunidad. Este enfoque tiene el potencial de personalizar las regulaciones y requisitos éticos, lo que puede ser crucial para abordar desigualdades particulares o sesgos que afectan a una población determinada. La localización permitiría ajustar los sistemas de IA para que respeten las particularidades geográficas, culturales y económicas de una región. No obstante, un enfoque puramente localizado correría el riesgo de entrar en conflicto con los valores fundamentales europeos, como el mercado común y el derecho a la igualdad. Si cada país o región estableciera sus propios estándares, podría socavarse la uniformidad normativa y generar fragmentación, creando barreras para las empresas que desearan operar en múltiples jurisdicciones.

permitiendo la participación de comunidades diversas y garantizando que no existan sesgos o discriminaciones en los sistemas.

> «Para hacer realidad la IA fiable, es preciso garantizar la inclusión y la diversidad a lo largo de todo el ciclo de vida de los sistemas de inteligencia artificial. Además de tener en cuenta a todos los afectados y garantizar su participación en todo el proceso, también es necesario garantizar la igualdad de acceso mediante procesos de diseño inclusivos, sin olvidar la igualdad de trato. Este requisito está estrechamente relacionado con el *principio de equidad*», (Grupo Independiente de Expertos, 2019, *Directrices éticas*, 79, p. 23) [39].

2.2.4. Sostenibilidad

La IA debe desarrollarse de manera sostenible, asegurando que no comprometa los recursos del planeta ni contribuya al deterioro del medio ambiente. El impacto ambiental de los sistemas de IA debe ser minimizado, optimizando la eficiencia energética y reduciendo los residuos tecnológicos durante todo su ciclo de vida. Ejemplo:

> «*Una IA sostenible y respetuosa con el medio ambiente.* Los sistemas de inteligencia artificial prometen ayudar a abordar algunas de las preocupaciones sociales más urgentes; no obstante, se debe garantizar que lo hagan del modo más respetuoso posible con el medio ambiente. En ese sentido, debería evaluarse en su integridad el proceso de desarrollo, despliegue y utilización de sistemas de IA, así como toda su cadena de suministro, a través, por ejemplo, de un examen crítico del

[39] Otras muestras:

- UNESCO: «Garantizar la diversidad y la inclusión. 19. El respeto, la protección y la promoción de la diversidad y la inclusión deberían garantizarse a lo largo del ciclo de vida de los sistemas de IA, de conformidad con el derecho internacional, en particular el derecho de los derechos humanos. Para ello se podría promover la participación activa de todas las personas o grupos, con independencia de su raza, color, ascendencia, género, edad, idioma, religión, opiniones políticas, origen nacional, étnico o social, condición económica o social de nacimiento, discapacidad o cualquier otro motivo» UNESCO, 2022, p. 8.

- Declaración de Montreal: «*AIS development and use must not lead to the homogenization of society through the standardization of behavior and opinions*», IVADO, 2018, Principio 7: Diversidad e Inclusión, p. 14.

- OECD: «1.1. *Inclusive growth, sustainable development and well-being. Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, well-being, sustainable development and environmental sustainability*», OECD, 2019, p. 8.

uso de los recursos y del consumo de energía durante la formación, dando prioridad a las opciones menos perjudiciales. Se deberían promover medidas que garanticen el respeto del medio ambiente por parte de todos los eslabones de la cadena de suministro», (Grupo Independiente de Expertos, 2019, *Directrices éticas*, 84, p. 24) [40].

2.2.5. Transparencia y Responsabilidad

A) Transparencia y Explicabilidad

Los sistemas de IA deben ser transparentes, lo que significa que su funcionamiento debe ser explicable y verificable por terceros, permitiendo auditorías que respalden su uso seguro y responsable.

La transparencia y la explicabilidad [41] están unidos [42] pues se refieren a la capacidad de los sistemas de IA para proporcionar explicaciones comprensibles sobre cómo y por qué se toman ciertas decisiones. La explicabilidad es crucial para la transparencia y la rendición de cuentas. Por ejemplo, en el ámbito financiero, los algoritmos que determinan la aprobación de préstamos deben poder explicar claramente los factores que influyeron en cada decisión.

---

[40] Otras citas:

- UNESCO: «Sostenibilidad. 31. El desarrollo de sociedades sostenibles depende del logro de un complejo conjunto de objetivos relacionados con distintas dimensiones humanas, sociales, culturales, económicas y ambientales. La llegada de las tecnologías de la IA puede beneficiar los objetivos de sostenibilidad o dificultar su consecución, dependiendo de la forma en que se apliquen en países con diferentes niveles de desarrollo. Por consiguiente, la evaluación continua de los efectos humanos, sociales, culturales, económicos y ambientales de las tecnologías de la IA debería llevarse a cabo con pleno conocimiento de las repercusiones de dichas tecnologías en la sostenibilidad como un conjunto de metas en constante evolución en toda una serie de dimensiones, como las que se definen actualmente en los Objetivos de Desarrollo Sostenible(ODS) de las Naciones Unidas», UNESCO, 2022, pp.10 y 11.

- Declaración de Montreal: «The development and use of AIS must be carried out so as to ensure a strong environmental sustainability of the planet», IVADO, 2018, Principio 10: Desarrollo Sostenible, p. 17.

[41] «*The addition of this principle, which we synthesise as 'explicability' both in the epistemological sense of 'intelligibility' (as an answer to the question 'how does it work?') and in the ethical sense of 'accountability' (as an answer to the question: 'who is responsible for the way it works?'), is therefore the crucial missing piece of the jigsaw when we seek to apply the framework of bioethics to the ethics of AI*», Floridi et al., 2018, p. 700.

[42] La explicabilidad está directamente relacionada con la transparencia y es uno de los campos de batalla desde el punto de vista ético. Por ejemplo O'Neil en su libro «*Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy"* aborda cómo los algoritmos y el big data pueden perpetuar y exacerbar las desigualdades sociales y llegar a amenazar la democracia y señala con claridad*: "To protect ourselves, we need to create a regulatory structure for Big Data and algorithmic decisions, one that ensures fairness and transparency*», O'Neil, 2016.

Ejemplos:

> «4. Transparencia 75) Este requisito guarda una relación estrecha con el *principio de explicabilidad* e incluye la transparencia de los elementos pertinentes para un sistema de IA: los datos, el sistema y los modelos de negocio, Grupo Independiente de Expertos, 2019, *Directrices éticas*, 75, p. 22.
>
> «77) *Explicabilidad*. La explicabilidad concierne a la capacidad de explicar tanto los procesos técnicos de un sistema de IA como las decisiones humanas asociadas (por ejemplo, las áreas de aplicación de un sistema de IA). La explicabilidad técnica requiere que las decisiones que adopte un sistema de IA sean comprensibles para los seres humanos y estos tengan la posibilidad de rastrearlas. Además, puede que sea necesario buscar un equilibrio entre la mejora de la explicabilidad de un sistema (que puede reducir su precisión) o una mayor precisión de este (a costa de la explicabilidad). Cuando un sistema de IA tenga un impacto significativo en la vida de las personas, debería ser posible reclamar una explicación adecuada del proceso de toma de decisiones del sistema de IA. Dicha explicación debería ser oportuna y adaptarse al nivel de especialización de la parte interesada (que puede ser una persona no experta en la materia, un regulador o un investigador). Además, debería ser posible disponer de explicaciones sobre la medida en que el sistema de IA condiciona e influye en el proceso de toma de decisiones de la organización, sobre las decisiones de diseño del sistema y sobre la lógica subyacente a su despliegue (garantizando así la transparencia del modelo de negocio)», Grupo Independiente de Expertos, 2019, *Directrices éticas*, 77, p. 22[43].

---

[43] Otras formulaciones:

- UNESCO: «Transparencia y explicabilidad 37. La transparencia y la explicabilidad de los sistemas de IA suelen ser condiciones previas fundamentales para garantizar el respeto, la protección y la promoción de los derechos humanos, las libertades fundamentales y los principios éticos. La transparencia es necesaria para que los regímenes nacionales e internacionales pertinentes en materia de responsabilidad funcionen eficazmente. La falta de transparencia también podría mermar la posibilidad de impugnar eficazmente las decisiones basadas en resultados producidos por los sistemas de IA y, por lo tanto, podría vulnerar el derecho a un juicio imparcial y a un recurso efectivo, y limita los ámbitos en los que estos sistemas pueden utilizarse legalmente» UNESCO, 2022, pp. 11 y 12.

- IEEE: «*Automated systems should generate audit trails recording the facts and law supporting decisions and such systems should be amenable to third-party verification to show that the trails reflect what the system in fact did. Audit trails should include a comprehensive history of decisions made in a case, including the identity of individuals who recorded the facts and their assessment of those facts. Audit trails should detail the rules applied in every mini-decision made by the system*», IEEE, 2019, p. 153.

- «Los conjuntos de datos y los procesos que dan lugar a la decisión del sistema de IA, incluidos los relativos a la recopilación y etiquetado de los datos así como a los algoritmos utilizados, deberían

B) Participación democrática

Este principio resalta la importancia de someter el desarrollo y uso de la IA al escrutinio democrático, promoviendo debates abiertos para garantizar que los sistemas sean justos y accesibles, por un lado, y, por otro, la IA no debe socavar los sistemas democráticos. Ejemplo:

> «*Respeto de la democracia, la justicia y el estado de Derecho*. En las democracias constitucionales, todo poder gubernamental debe estar autorizado legalmente y limitado por la legislación. Los sistemas de IA deberían servir para mantener e impulsar procesos democráticos, así como para respetar la pluralidad de valores y elecciones vitales de las personas. Los sistemas de IA no deben socavar los procesos democráticos, las deliberaciones humanas ni los sistemas democráticos de votación. Asimismo, los sistemas de IA deben incluir un compromiso de garantizar que su funcionamiento no menoscabe los compromisos esenciales en los que se fundamenta el estado de Derecho -así como las leyes y reglamentos de obligado cumplimiento- y de asegurar el respeto de las garantías procesales y la igualdad ante la ley», (Grupo Independiente de Expertos, 2019, *Directrices éticas*, 43, p. 13) [44].

C) Prevención del daño Rendición de cuentas y Responsabilidad

Aunque la IA pueda tomar decisiones, la responsabilidad final de las acciones derivadas de estas decisiones debe recaer siempre en los seres huma-

---

documentarse con arreglo a la norma más rigurosa posible con el fin de posibilitar la trazabilidad y aumentar la transparencia, Grupo Independiente de Expertos, 2019, *Directrices éticas*, p. 22).

- OECD: «1.3. *Transparency and explainability. AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:*

i. to foster a general understanding of AI systems, including their capabilities and limitations,

ii. to make stakeholders aware of their interactions with AI systems, including in the workplace,

iii. where feasible and useful, to provide plain and easy-to-understand information on the sources of data/input, factors, processes and/or logic that led to the prediction, content, recommendation or decision, to enable those affected by an AI system to understand the output, and,

iv. to provide information that enable those adversely affected by an AI system to challenge its output», OECD, 2019, pp. 8 y 9.

- Principios Asilomar «7) *Failure Transparency: If an AI system causes harm, it should be possible to ascertain why*»

- «8) *Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority*» Future of Life Institute, 2017.

[44] Otra version: «*AIS must meet intelligibility, justifiability, and accessibility criteria, and must be subjected to democratic scrutiny, debate, and control*», (IVADO, 2018, Principio 5: Participación Democrática, p. 12).

nos. Los desarrolladores y operadores de IA deben ser responsables y rendir cuentas de los sistemas que crean, asegurando que puedan responder por sus efectos y decisiones. Además debe existir un sistema eficaz de responsabilidad para indemnizar a las víctimas por los daños que los sistemas de IA pudieran causar. Hay que recordar que no dañar es un principio ético esencial. Ya lo era en las Leyes de la Robótica[45] que imaginó Isaac Asimov, y que son citadas expresamente por las Normas de Derecho Civil sobre Robótica[46].

Ejemplos:

> «7. Rendición de cuentas 87) Los requisitos anteriores se complementan con el de rendición de cuentas, estrechamente relacionado con el *principio de equidad*. Este requisito exige establecer mecanismos que permitan garantizar la responsabilidad y rendición de cuentas sobre los sistemas de IA y sus resultados, tanto antes de su implantación como después de esta» Grupo Independiente de Expertos, 2019, *Directrices éticas*, 87, p. 22.
>
> «91) *Compensaciones*. Cuando se produzcan efectos adversos injustos, deberían preverse mecanismos accesibles que aseguren una compensación adecuada. El hecho de saber que se podrá obtener una reparación si las cosas no salen según lo previsto es crucial para garantizar la confianza. Se debería prestar atención a las personas o grupos vulnerables», Grupo Independiente de Expertos, 2019, *Directrices éticas*, 91[47].

---

[45] «*1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws*».

[46] Parlamento Europeo, 2017, *Normas de Derecho civil*.

[47] Otras explicaciones:

- UNESCO: «68. Los Estados Miembros deberían elaborar, examinar y adaptar, según proceda, marcos reguladores para alcanzar la rendición de cuentas y la responsabilidad por el contenido y los resultados de los sistemas de IA en las diferentes etapas de su ciclo de vida. Cuando sea necesario, los Estados Miembros deberían introducir marcos de responsabilidad o aclarar la interpretación de los marcos existentes para garantizar la atribución de la responsabilidad por los resultados y el funcionamiento de los sistemas de IA. Además, al elaborar los marcos reguladores, los Estados Miembros deberían tener en cuenta, en particular, que la responsabilidad y la rendición de cuentas deben recaer siempre en última instancia en personas físicas o jurídicas y que no se debe otorgar personalidad jurídica a los propios sistemas de IA. Para lograrlo, esos marcos reguladores deberían ajustarse al principio de la supervisión humana y establecer un enfoque global centrado en los actores de la IA y los procesos tecnológicos que intervienen en las diferentes etapas del ciclo de vida de los sistemas de IA», UNESCO, 2022, 68, pp. 17 y 18.

- Declaración de Montreal: «*Only human beings can be held responsible for decisions stemming from recommendations made by AIS, and the actions that proceed therefrom*», IVADO, 2018, Principio 9: Responsabilidad, p. 16.

3. INTENTO DE APORTAR SOLUCIONES A LOS DESAFÍOS A LOS QUE SE ENFRENTA LA RC POR DAÑOS CAUSADOS POR LA IA

La RC en el contexto de la IA plantea varios desafíos cuando se consideran los principios y documentos éticos que hemos revisado. A continuación se identifican algunos de los más importantes.

En las siguientes páginas se ofece una opinión sobre la posible solución a los mismos desde los principios éticos analizados. Nótese que los aportes realizados se basan en los principios éticos, con la intención de enriquecer el debate, sin entrar en otros aspectos.

3.1. RC objetiva o subjetiva

Una de las cuestiones que siempre aparece en primer lugar cuando se trata sobre la RC en materia de IA es cuál debe ser su naturaleza: objetiva o subjetiva[48].

Los documentos y principios éticos revisados proporcionan orientación sobre cómo enfrentar esta cuestión, pero no ofrecen una respuesta directa sobre si la RC debe ser subjetiva u objetiva. Aun así, ayudan a orientar la decisión.

En este sentido y a sabiendas de que existen muchísimos más factores que pueden influir en la decisión que se adopte (factores económicos, sociológicos, de mercado, geoestratégicos…), a partir de los principios éticos señalados anteriormente, se pueden realizar una serie de consideraciones:

- OECD: «1.5. *Accountability a) AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of the art.*

b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outputs and responses to inquiry, appropriate to the context and consistent with the state of the art.

c) AI actors, should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on an ongoing basis and adopt responsible business conduct to address risks related to AI systems, including, as appropriate, via co-operation between different AI actors, suppliers of AI knowledge and AI resources, AI system users, and other stakeholders. Risks include those related to harmful bias, human rights including safety, security, and privacy, as well as labour and intellectual property rights», OECD, 2019, p. 9.

- IEEE: «*Lawmakers and enforcers need to ensure that the implementation of A/IS is not abused by businesses and entities employing the A/ IS to avoid liability or payment of damages. Governments should consider adopting regulations requiring insurance or other guarantees of financial responsibility so that victims can recover damages for harm that A/ IS cause*», IEEE, 2019, p. 155.

- «*Objective: Provide effective regulation of A/IS to ensure public safety and responsibility while fostering a robust AI industry*», IEEE, 2019, p. 188.

[48] Velasco Perdigones, 2021.

1) La insistencia de los documentos en la **transparencia y explicabilidad** exige que la IA fiable y éticamente buena tenga la «capacidad de explicar tanto los procesos técnicos de un sistema de IA como las decisiones humanas asociadas (por ejemplo, las áreas de aplicación de un sistema de IA)»[49]. Esta repetición de la transparencia podría hacernos pensar en una **RC subjetiva,** pues si fuera posible rastrear y explicar de dónde procede el daño que la IA ha causado, lo lógico sería atribuir la culpa y la RC al actor humano responsable de la acción u omisión que ha ocasionado el daño (que podría ser el fabricante, el diseñador, el operador…).

2) De forma parecida los principios de **autonomía humana** y de control humano siempre conducen a que exista una responsabilidad humana (recordemos ideas como: «*Only human beings can be held responsible for decisions stemming from recommendations made by AIS, and the actions that proceed therefrom*»[50]. Si en el tipo de IA que queremos siempre hay una persona que es responsable, parecería lógico entender que ese responsable lo fuera también civilmente, lo que podría encajar mejor con una **RC subjetiva**.

3) Diferente resultado podría arrojar el análisis si se consideraran otros otros principios como los de **seguridad y prudencia, solidez y seguridad técnica**. El enfoque en la prevención y en la robustez técnica nos haría pensar que, aunque no se pudiera identificar la culpabilidad de un actor humano (lo que supondría que han fracasado los principios anteriormente citados), las víctimas deberían tener derecho a una compensación por los daños que hubieran sufrido, lo que nos llevaría a una **RC objetiva** pues se indemnizaría con independencia de poder atribuir una negligencia a un actuar humano.

4) Algo parecido ocurre si se tuvieran en cuenta los principios de **prudencia** o la **prevención del daño**. Es verdad que estos principios guían la actuación ex ante, pues buscan anticipar posibles consecuencias negativas y mitigar el mal uso, lo que podría llevar aparejado que debería exonerarse de responsabilidad a quien ha actuado con prudencia y prevención extremando su deber de cuidado, pero por otro lado, ante daños que son de naturaleza impredecible (fuera de la posible prevención), resultaría injusto exigir la prueba de la negligencia y la **RC objetiva** sería una solución más pragmática.

---

[49] Grupo Independiente de Expertos, 2019, *Directrices éticas*, 77, p. 22.

[50] IVADO, 2018, Principio 9: Responsabilidad, p. 16.

La RC subjetiva requiere la prueba de la culpa del causante del daño: la acción u omisión de la que trae causa el daño (que la realizó el fabricante, desarrollador, operador, usuario…) es negligente. Para poder operar en esta materia se necesita que el mal funcionamiento o uso indebido de la IA pueda ser explicado y atribuido a errores humanos. Sin embargo, tal nivel de explicabilidad no siempre es posible en sistemas de *deep learning* donde el comportamiento de los algoritmos puede presentarse de manera opaca y por ello resulte imposible averiguar y probar la culpa.

Hasta el concepto de culpa se está viendo trastocado con la IA[51]. Cuando una IA se equivoca no es fácil saber si ha sido negligente, porque no existe un estandar de diligencia que les sea aplicable[52]. Si se piensa en la diligencia de la persona que usa la IA, también hay cambios porque ahora una persona razonable como estándar de diligencia será una persona que usa IA[53].

Por su parte, para la RC objetiva basta con probar el daño y que este deriva causalmente de la actividad de la IA, aunque no se pueda precisar con una concreción exhaustiva el momento exacto del que deriva el comportamiento dañoso posterior. Por esta razón, la RC objetiva parece más aconsejable de acuerdo con los principios de seguridad y prevención del daño pues facilita la compensación a las víctimas que han sufrido daños, sobre todo si los principio de transparencia y explicabilidad no se han cumplido todo lo que debían.

De resultas de lo anterior, y aun siendo cuestionable, considero que la RC objetiva es más acorde con todos los textos que se refieren a los principio éticos de la IA. En todos ellos es habitual señalar la responsabilidad de los desarrolladores y operadores de la IA, aunque sea difícil la prueba de la negligencia en sistemas tecnológicos tan complejos. La RC objetiva sería coherente con los principios de equidad, seguridad y prevención del daño entre otros. La RC subjetiva, por su parte, podría servir para azuzar la búsqueda de la di-

---

[51] Los retos que la IA supone para el concepto de negligencia se pueden ver en Selbst, 2020.

[52] Como señala Borghetti: «*When one tries to assess the existence of a human fault or negligence, one always compares the behavior of the defendant with the one which a model human being would have adopted in the same circumstances. Such a comparison is valid and pertinent because we assume that all human beings share the same kind of rationality and should be able to figure out what is reasonable in any kind of circumstances. The problem is that algorithms do not function the way human beings do, and the outcomes they produce may not be the product of a human-like rationality. Besides, two algorithms designed to perform the same tasks might function along different types of rationality and may therefore face one same situation in very different ways*», Borghetti, 2019, 94-102.

[53] «*For example, once self-driving cars become safer than traditional vehicles, a jury might find that it is unreasonable to drive yourself rather than to use a self-driving car. Applying the "reasonable person using an autonomous computer" standard to the earlier hypothetical involving a child running into the street, the human driver's negligence would not be based on failing to stop in 100 feet as a self-driving car would have; rather, liability would be based on her driving in the first place. A reasonable person would not have driven*», Abbott, 2018, 37.

ligencia en todos los actores y para hacer responsables a los negligentes en aquellos casos en los que se pueda probar[54].

En cualquier caso, saber si conviene una RC objetiva o subjetiva es una duda que actualmente está instalada en los legisladores europeos. El art. 5 de la Propuesta de Directiva sobre responsabilidad en materia de IA impone que «la Comisión revisará la aplicación de la presente Directiva y presentará un informe al Parlamento Europeo, al Consejo y al Comité Económico y Social Europeo, acompañado, en su caso, de una propuesta legislativa». En dicho informe «En particular, deberá evaluar la idoneidad de las normas de responsabilidad objetiva (sin culpa) para las demandas contra los operadores de determinados sistemas de IA -siempre que estas no estén ya reguladas por otras normas de responsabilidad de la Unión- y la necesidad de aseguramiento, teniendo en cuenta al mismo tiempo el efecto y el impacto en la introducción general y la adopción de los sistemas de IA, especialmente para las pymes»[55].

3. 2. La Explicabilidad y Opacidad de los Sistemas de IA (Black Box)

Uno de los desafíos más significativos de la RC en esta materia es la opacidad o la falta de explicabilidad en los sistemas de IA, especialmente en tecnologías como las que incorporan procesos de *deep learning*. Los sistemas de IA pueden tomar decisiones basadas en complejos algoritmos que no siempre son comprensibles para los seres humanos. En estos casos, si el sistema de IA actúa de manera autónoma y no se puede explicar cómo llegó a una decisión o a causar un daño[56], puede ser extremadamente difícil identificar a

---

[54] Desde un punto de vista fundamentalmente jurídico véanse los argumentos muy interesantes y proclives a la RC subjetiva que propone Atienza Navarro, 2023, 12, 3.2 y 3.3.

[55] Conviene también recordar la Propuesta de Reglamento del Parlamento Europeo y del Consejo relativo a la responsabilidad civil por el funcionamiento de los sistemas de IA de 2020, en cuyo Considerando 14 se decía «Reconoce que el tipo de sistema de IA sobre el que el operador ejerce control es un factor determinante en lo que respecta a la responsabilidad; observa que un sistema de IA que conlleve un alto riesgo inherente y actúe de manera autónoma potencialmente pone en peligro en mucha mayor medida al público en general; considera que, habida cuenta de los retos jurídicos que plantean los sistemas de IA para los regímenes de responsabilidad civil existentes, parece razonable establecer un régimen común de responsabilidad objetiva para los sistemas de IA autónomos de alto riesgo; subraya que este enfoque basado en el riesgo, que puede incluir varios niveles de riesgo, debe fundamentarse en criterios claros y una definición adecuada de riesgo elevado, así como ofrecer seguridad jurídica». Efectivamente, en el art. 4 se establecía una RC objetiva para los sistemas de IA de alto riesgo y en el art. 8 una RC subjetiva par los demás sistemas de IA.

[56] Benhamou y Ferland explican este efecto: «*AI's designers do not program all the possible scenarios in advance, nor give specific instructions for each of them; rather, they set a goal for the machine and let AI process the data input, learn from it, and decide the best course of action to reach its goal. This leads to the scenario where the AI's programmers may not have exact understanding of how it reached such goal or what the stages leading to*

un responsable o determinar si hubo negligencia. No se trata de algo aislado, sino de algo común que suele denominarse como el efecto "caja negra"[57]. La opacidad es un obstáculo tanto para la RC subjetiva, como para la RC objetiva, porque afecta al nexo causal imprescindible en ambas para atribuir la RC[58].

La opacidad afecta además a la investigación y desarrollo de la IA, lo que también pone en entredicho el principio de participación democrática[59].

Es verdad que una de las exigencias éticas expuesta en los documentos analizados es la transparencia y explicabilidad, pero es que ni siquiera es técnicamente posible en todas las aplicaciones de la IA. Por otro lado, es también común afirmar que la responsabilidad siempre debe recaer en persona físicas o jurídicas que sean legalmente responsables[60]. No podemos olvidar que des-

---

success were; in other words, they cannot explain the AI's "thought process" leading to the final result. The same is true for AI's failures which cannot always be explained or understood by humans. For instance, algorithms in precision medicine process patient and hospital data to predict patient risks and formulate diagnoses, but it is not always possible to identify which data elements were processed, the weight that was given to each element in the global assessment and whether there are unethical bias in the processing. Even in cases where the algorithm itself is rather simple, the data fed into the algorithm may be so diverse and ever changing (in the case of autonomous vehicles, one may think about inputs from cameras, sensors, lasers, microphones, etc.) that it is often impossible to reproduce the environment in which the injury happened and identify its source», Benhamou y Ferland, 2021, 8.

[57] «Que una aplicación que se apoye en la IA actúe de manera autónoma quiere decir que lleva a cabo una tarea sin que cada paso esté predefinido y que lo hace con menos o, en última instancia, sin ningún control o supervisión humanos inmediatos. Los algoritmos basados en el aprendizaje automático de la máquina pueden ser difíciles, si no imposible, de comprender («efecto caja negra»)», Comisión Europea, 2020, *Informe sobre las repercusiones*, p. 18.

[58] «La transparencia es necesaria para que los regímenes nacionales e internacionales pertinentes en materia de responsabilidad funcionen eficazmente» (UNESCO, 2022, 37).

[59] «*From a regulatory standpoint, some of the most problematic features of AI are not features of AI itself, but rather the manner in which AI R&D work can be done. Discreetness refers to the fact that AI development work can be conducted with limited visible infrastructure. Diffuseness means that the individuals working on a single component of an AI system might be located far away from one another. A closely related feature, discreteness, refers to the fact that the separate components of an AI system could be designed in different places and at different times without any conscious coordination. Finally, opacity denotes the possibility that the inner workings of an AI system may be kept secret and may not be susceptible to reverse engineering. Each of these features is shared, to varying degrees, by R&D work on many technologies in the Information Age, but they present particularly unique challenges in the context of AI*», Scherer, 2016, 370.

[60] Véase por ejemplo: «Los Estados Miembros deberían velar por que siempre sea posible atribuir la responsabilidad ética y jurídica, en cualquier etapa del ciclo de vida de los sistemas de IA, así como en los casos de recurso relacionados con sistemas de IA, a personas físicas o a entidades jurídicas existentes. La supervisión humana se refiere, por tanto, no solo a la supervisión humana individual, sino también a la supervisión publica inclusiva, según corresponda» (UNESCO, 2022, p. 35).

«Los Estados Miembros deberían elaborar, examinar y adaptar, según proceda, marcos reguladores para alcanzar la rendición de cuentas y la responsabilidad por el contenido y los resultados de los sistemas de IA en las diferentes etapas de su ciclo de vida. Cuando sea necesario, los Estados Miembros deberían introducir marcos de responsabilidad o aclarar la interpretación de los marcos existentes para garantizar la atribución de la responsabilidad por los resultados y el funcionamiento de los sistemas de IA. Además, al elaborar los marcos reguladores, los Estados Miembros deberían tener en cuenta, en particular, que la responsabilidad y la rendición de cuentas deben recaer siempre

de un punto de vista jurídico la imprevisibilidad de la IA ha sido programada y por tanto prevista por los diseñadores[61].

En lo que toca a la explicabilidad, no se trata de optar entre una solución u otra, sino de buscar posibles respuestas para reducir o intentar eliminar la opacidad. Diversas posibilidades, no exhaustivas, pueden ser:

1. **Reforzar la trazabilidad y la explicabilidad**. Los propios sistemas de IA deberían incorporar mecanismos de trazabilidad y explicabilidad que permitieran rastrear las decisiones de la IA. Se trataría de procedimientos parecidos a lo que se denomina en el mundo de la aviación como "caja negra" y que es un registro de vuelo en el que se almacenan datos importantes de lo que acontece durante el vuelo para poder averiguar posteriormente lo que ha pasado.

Conviene recordar que el principio de transparencia está directamente relacionado con la prevención del daño, la rendición de cuentas y la responsabilidad. Mecanismos que favorecieran la trazabilidad y la explicabilidad favorecerían la audatibilidad de tal manera que sería más comprensible la comprensión del comportamiento de la IA para los usuarios, para los reguladores y para todas les personas afectadas. Sin duda también favorecería la prueba de la relación causal en los procedimientos de RC.

2. Desarrollar **herramientas de certificación y auditorías independientes**. Además de los mecanismo internos de la IA que favorezcan la explicabilidad, sería conveniente el desarrollo de auditorías independientes que permitieran analizar y verificar el comportamiento de los sistemas de IA.

Antes de comercializar o implementar un sistema de IA, este debería pasar por un proceso de **certificación** que asegurara que cumple con los estándares de seguridad, transparencia y explicabilidad establecidos en los principios éticos[62].

---

en última instancia en personas físicas o jurídicas y que no se debe otorgar personalidad jurídica a los propios sistemas de IA. Para lograrlo, esos marcos reguladores deberían ajustarse al principio de la supervisión humana y establecer un enfoque global centrado en los actores de la IA y los procesos tecnológicos que intervienen en las diferentes etapas del ciclo de vida de los sistemas de IA» (UNESCO, 2022, p. 68).

[61] «*Thus, a learning AI's designers will not be able to foresee how it will act after it is sent out into the world - but again, such unforeseeable behavior was intended by the AI's designers, even if a specific unforeseen act was not*», Scherer, 2016, pp. 36-37.

[62] Señalan las Directrices éticas: «Dado que no cabe esperar que todas y cada una de las personas comprendan plenamente el funcionamiento y los efectos de los sistemas de IA, puede tenerse en consideración a aquellas organizaciones que puedan acreditar ante el público que un sistema de IA es transparente, responsable y equitativo. Estas certificaciones aplicarían normas desarrolladas para diferentes ámbitos de aplicación y técnicas de IA, convenientemente alineadas con las normas industriales y sociales del contexto específico de que se trate. », Grupo Independiente de Expertos, 2019, *Directrices éticas*, 107.

Los sistemas de IA deberían estar sujetos a auditorías periódicas que verificaran su funcionamiento y permitieran la trazabilidad de las decisiones que toman. Si se produce un daño, la auditoría permitiría identificar en qué punto del proceso ocurrió el fallo, facilitando así la determinación de responsabilidad[63].

En la misma línea podrían explorarse sistemas de certificación y auditoría que incluyeran estándares técnicos que garantizaran que los sistemas de IA puedan ser comprendidos y cumplen con los principios de transparencia y explicabilidad.

3. Considerar la **RC objetiva** mitigaría parcialmente los efectos negativos de la opacidad. Al no exigir la prueba de la negligencia y bastar con demostrar la relación de causalidad, la opacidad tendría menos campo de juego. Esto alinearía la RC objetiva con los principios de seguridad, prudencia, rendición de cuentas y garantizaría mejor la compensación de las víctimas que también es una idea que aparece, por ejemplo, en las Directrices éticas (Grupo Independiente de Expertos, 2019, *Directrices éticas*, 91, p. 25).

4. **Introducir la obligación de exhibición de prueba y presunciones**. La Propuesta de Directiva sobre responsabilidad en materia de IA establece normas sobre la exhibición de pruebas en los procesos de RCE por los daños causados por sistemas de IA de alto riesgo y sobre la carga de prueba en la RCE subjetiva por daños causados por sistemas de IA. Una de las razones importantes para introducir estos mecanismos es intentar contrarrestar el efecto caja negra:

> «Las características específicas de la IA, incluidas su complejidad, su autonomía y su opacidad (el denominado efecto de «caja negra»), pueden dificultar o hacer excesivamente costoso para las víctimas determinar cuál es la persona responsable y probar que se cumplen los requisitos para una demanda de responsabilidad civil admisible. En particular, al reclamar una indemnización, las víctimas podrían tener que soportar unos costes iniciales muy elevados y enfrentarse a procedimientos judiciales mucho más largos, en comparación con los casos sin relación alguna con la inteligencia artificial. Por lo tanto, las víctimas

[63] «La auditabilidad es la capacidad para evaluar los algoritmos, los datos y los procesos de diseño.

Esto no implica necesariamente que siempre deba disponerse de forma inmediata de la información sobre los modelos de negocio y la propiedad intelectual del sistema de IA. La evaluación por parte de auditores internos y externos y la disponibilidad de los correspondientes informes de evaluación pueden contribuir a la fiabilidad de esta tecnología. En aplicaciones que afecten a los derechos fundamentales, incluidas las aplicaciones esenciales desde el punto de vista de la seguridad, los sistemas de IA deberían poder someterse a auditorías independientes», Grupo Independiente de Expertos, 2019, *Directrices éticas*, 88.

> pueden verse disuadidas de intentar siquiera obtener una indemnización», Exposición de Motivos de la Propuesta de Directiva sobre responsabilidad en materia de IA.

Tanto la exhibición de prueba obligatoria, como las presunciones de causalidad son herramientas que están acordes con los principios de prevención del daño y responsabilidad para los supuestos en los que la transparencia y la explicabilidad no se han podido alcanzar.

5. **Fomentar la supervisión humana**. La supervisión humana es un principio que aparece en numerosos documentos y precisamente coadyuvaría a facilitar la transparencia y explicabilidad de la IA. En la medida en que un sistema de IA este sometido a un control humano y exista una supervisión humana para las decisiones más relevantes, el proceso sería más explicable, sobre todo si esta supervisión se exige ante las decisiones más críticas.

3. 3. Difusión de la Responsabilidad

El desarrollo y uso de IA involucra múltiples actores: diseñadores, desarrolladores de software, fabricantes de hardware, operadores, usuarios... Esto provoca una difusión de la responsabilidad entre las distintas partes, complicando la identificación de quién debe responder en caso de que se cause un daño[64].

Como señala Wagner (2018, pp. 9-10), también es importante considerar la dispersión del control resultante del desacoplamiento de las actividades. Un producto o hardware puede soportar diferentes programas de distintos programadores que los usuarios pueden descargar y modificar a voluntad. Se puede combinar el software original con otro e ir variándolo hasta que quede prácticamente en una combinación única, como ocurre con muchos de nuestros teléfonos móviles. También se pueden juntar varias piezas de hardware u otros tipos de hardware, lo que hace que la tarea de atribuir responsabilidades sea muy complicada de acuerdo con los esquemas que tenemos hoy en día.

El desafío que esto representa es tan simple como: ¿quién debe ser responsable, el fabricante, el desarrollador del algoritmo, el proveedor del hardware, el usuario final que opera el sistema…? La falta de claridad en este punto puede complicar mucho las demandas de RC.

---

[64] Como señalan Benhamou y Ferland «*The number of stakeholders involved in the creation and operation of AI systems is concurrently rising: hardware manufacturers, software designers, sellers, equipment and software installers, facility owners, AI owners, AI users and trusted third parties, amongst others, may all have a role to play in ensuring that AI does not cause harm, and allocating liability in this context is not an easy task*», Benhamou y Ferland, 2021, pp. 168-170.

El principio de rendición de cuentas puede verse afectado si no hay claridad en este asunto, pues recordemos que «Este requisito exige establecer mecanismos que permitan garantizar la responsabilidad y rendición de cuentas sobre los sistemas de IA y sus resultados, tanto antes de su implantación como después de esta»[65].

Las posibles respuestas ante este desafío pueden ser variadas:

1. **Responsabilidad compartida en función del rol y control.** Como señala el documento de la OECD (2019, p. 9): «*AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of the art*».

   Así se podría distribuir la responsabilidad entre los diversos actores en función del grado de control[66] y los diversos roles que cada uno tiene, de tal forma que habría una responsabilidad compartida[67]. Sin pretender ser exhaustivo, podrían existir distintas responsabilidades:

   — **Desarrollador del algoritmo**: sería responsable en la medida en que hubiera fallos en el diseño del algoritmo o se hubieran ignorado medidas de precaución razonables para evitar sesgos, errores o inseguridades. Este enfoque se basa en el principio de precaución y responsabilidad. Además los desarrolladores serían unos de los principales responsables de la robustez y seguridad técnica de la IA que también aparece en diversos documentos.

---

[65] Grupo Independiente de Expertos, 2019, *Directrices éticas*, 87, p. 22.

[66] Conviene recordar el Libro Blanco sobre la inteligencia artificial - un enfoque europeo orientado a la excelencia y la confianza: «La Comisión considera que, en un futuro marco regulador, cada obligación debe dirigirse a la(s) persona(s) que esté(n) en mejor posición para abordar todo posible riesgo. Por ejemplo, mientras que los desarrolladores de IA pueden ser los que estén en mejor posición para abordar los riesgos derivados de la fase de desarrollo, su capacidad de controlar los riesgos durante la fase de uso puede ser más limitada. En este caso, el implementador debe ser objeto de la obligación correspondiente. Ello debe entenderse sin perjuicio de determinar qué parte debe ser responsable de los daños causados, a efectos de la responsabilidad civil ante los usuarios finales u otras partes que sufran daños, y de ofrecer un acceso efectivo a la justicia. Con arreglo a la legislación de la UE sobre responsabilidad con relación a los productos, la responsabilidad civil por los productos defectuosos se atribuye al productor, sin perjuicio de la legislación nacional, que también puede contemplar una indemnización a cargo de otras partes», Comisión Europea, 2020, *Libro blanco*, p. 27.

[67] En un artículo con un título muy descriptivo sobre la autonomía de la IA (*Machines without principals: Liability rules and artificial intelligence*) Vladeck proponía que, especialmente en casos donde las máquinas actúan de manera impredecible o cometen errores que no pueden ser atribuidos a fallos de diseño o fabricación, se estableciera una responsabilidad objetiva colectiva bajo la doctrina de la *«common enterprise» («each entity within a set of interrelated companies may be held jointly and severally liable for the actions of other entities that are part of the group»*). Se trataría de una RC cuyo fundamento último sería simplemente «*to protect a blameless party (the person who sustained injured) by making others bear the cost*», Vladeck, 2014, pp. 149 y ss.

— **Proveedor del hardware**: sería responsable si el fallo se originase en el hardware o si el hardware no fuera adecuado para las especificaciones del software o sistema de IA. Aquí se podría aplicar la RC objetiva para los defectos de fabricación, en línea con los principios de seguridad y robustez.

— **Fabricante**: la RC del fabricante se encuentra recogida en la Propuesta de Directiva sobre responsabilidad por los daños causados por productos defectuosos, incluyendo en su ámbito objetivos los productos que incorporan sistemas de IA. Se trata de una RC objetiva ampliamente aceptada por la doctrina, la industria y los consumidores.

— **Operador/usuario final**: si el usuario final operara el sistema de IA de manera negligente o incorrecta, por ejemplo, ignorando advertencias de seguridad o utilizando el sistema en un entorno no adecuado, debería asumir la responsabilidad. Los principios de transparencia y explicabilidad exigen que los usuarios sean conscientes de cómo operar y controlar la IA.

2. Podría pensarse en establecer un **sistema mixto de RC objetiva combinado con elementos subjetivos**. En la misma línea de lo anterior, se podría establecer una RC objetiva de todos los actores involucrados en el sistema de IA, de tal forma que la víctima fuera compensada de los daños y al mismo tiempo tomar en consideración la existencia de negligencia si se probase, para penalizar a los culpables. Esta matización subjetiva podría afectar, por ejemplo, al usuario final de la IA que haya sido negligente. Véase lo que se dice en el Principio 9 de la Declaración de Montreal: «5. *When damage or harm has been inflicted by an AIS, and the AIS is proven to be reliable and to have been used as intended, it is not reasonable to place blame on the people involved in its development or use*» (IVADO, 2018). A contrario, si se puede probar la negligencia del usuario final, lo lógico sería hacer responsable a este último.

3. **Fondo de compensación para víctimas de IA.** En aquellos casos en que sea imposible identificar a un responsable directo o cuando exista un daño difuso que afecte a múltiples personas de manera simultánea, podría crearse un fondo de compensación[68]. Este fondo sería financiado por las empresas que desarrollan, venden o implementan sistemas de IA y garantizaría que las víctimas recibieran compensación, independientemente de las dificultades para probar la culpa o negligencia.

[68] Por ejemplo, con ciertos matices, para el ámbito de los productos defectuosos propone un seguro obligatorio y un fondo de compensación García Teruel, 2021, 68, 2.4.

En la UE se está planteando esta posibilidad desde las primeras reflexiones[69] y lo volvieron a dejar abierto el grupo de expertos de alto nivel en su segundo entregable[70]. Esto sería conforme con diversos principios como los de compensación y equidad, pues evitarían abandonar a su suerte a las víctimas cuando no fuera posible identificar al responsable directo.

4. **Cláusulas contractuales de responsabilidad entre actores.** En los contratos entre los diversos actores involucrados en el ciclo de vida de los sistemas de IA sería aconsejable establecer cláusulas contractuales que concretaran los diversos roles y las responsabilidad de cada uno, lo que ayudaría a fijar una RC más adecuada, proporcionada y justa.

Igualmente, los contratos de uso de IA también deberían incluir disposiciones que aclararan las responsabilidades en caso de mal uso por parte del usuario o de fallos en la configuración o integración del sistema.

3.4. Autonomía de la IA y Responsabilidad Humana

La característica que diferencia la IA de cualquier tecnología precedente es la habilidad de la IA para actuar autónomamente[71]. A medida que la IA se vuelve más autónoma, surgen preguntas sobre hasta qué punto los humanos pueden o deben ser responsables de las decisiones que toman estos sistemas, especialmente si la IA actúa de manera impredecible o fuera del control de sus operadores[72]. Si bien los principios éticos subrayan que la **responsabilidad final siempre recae en los seres humanos**[73], es posible que los operadores

---

[69] Decían por ejemplo las Normas de Derecho civil sobre robótica: «Considera que, tal como sucede con el seguro de vehículos de motor, dicho sistema podría completarse con un fondo que garantizara la reparación de daños en los casos de ausencia de una cobertura de seguro; pide al sector de los seguros que desarrolle nuevos productos y tipos de ofertas adaptados a los progresos de la robótica», Parlamento Europeo, 2017, *Normas de Derecho civil*, 58.

[70] «*Finally, civil liability rules must be able to ensure adequate compensation in case of harm and/or rights violations (either through strict or tort liability), and may need to be complemented with mandatory insurance provisions*», Grupo Independiente de Expertos, 2019, *Policy and investment recommendations*, 27.2.

[71] Scherer, 2016, p. 364.

[72] Las Normas de Derecho civil sobre robótica consideraban que «en el supuesto de que un robot pueda tomar decisiones autónomas, las normas tradicionales no bastarán para generar responsabilidad jurídica por los daños ocasionados por el robot, ya que no permitirán determinar la parte que ha de hacerse cargo de la indemnización, ni exigir a dicha parte que repare el daño ocasionado», Parlamento Europeo, 2017, *Normas de Derecho civil*, AF.

[73] Esto ocurre habitualmente en Derecho. Por ejemplo Magrani recuerda que el art. 12 de la Convención de las Naciones Unidas sobre la Utilización de las Comunicaciones Electrónicas en los Contratos Internacionales supone que una persona es responsable en última instancia de la acción generada por una máquina, cuando establece: «*Article 12. Use of automated message systems for contract formation. A contract formed by the interaction of an automated message system and a natural person, or by the*

o diseñadores no puedan prever todas las acciones de la IA, lo que genera incertidumbre en cuanto a la RC[74].

La RC clásica comienza a resentirse cuando nos enfrentamos a sistemas complejos (compuestos por elementos diversos) en los que los cursos causales son difíciles de determinar y es difícil averiguar cuáles han sido las causas reales de los daños. En la IA, no se trata solo de complejidad, sino también de elementos interconectados y dependencia de datos externos, que le dan autonomía al sistema[75]. La complejidad engloba todo el entorno en el que un sistema interactúa con otros dispositivos, productos, servicios, etc., configurando un ecosistema complejo donde el ecosistema puede llegar a ser tan inmenso como todo Internet.

Además, el riesgo de un producto puede surgir de la conectividad del producto, como: una pulsera inteligente para niños que permite rastrear y contactar con el niño (notificación RAPEX islandesa publicada en el sitio web de EU Safety Gate (A12/0157/19) o un vehículo cuyo software presenta fallos de seguridad de tal manera que permite que un tercero no autorizado acceda al sistema de control (notificación RAPEX alemana publicada en el sitio web de EU Safety Gate (A12/1671/15)).

Las respuestas a esta cuestión también pueden hacerse a través de otras ya vistas en los apartados anteriores:

1. **RC objetiva, especialmente en los sistemas de IA altamente autónomos**. La RC objetiva parece más aconsejable en los casos de sistemas de IA autónomos, en los que quizás no es fácil descubrir una negligencia más allá de una genérica culpa al poner en funcionamiento una IA autónoma y por ello impredecible en algunos supuestos.

2. **Fondos de compensación.** Por la misma razón anterior, en estos casos de sistemas de IA altamente autónomos parece más aconsejable la creación de fondos de compensación para las víctimas de la IA.

3. **Certificación y auditoria de los sistemas de IA altamente autónomos**. Como quiera que la autonomía se acompaña de impredecibilidad, resultaría especialmente importante que este tipo de sistemas cons-

---

interaction of automated message systems, shall not be denied validity or enforceability on the sole ground that no natural person reviewed or intervened in each of the individual actions carried out by the automated message systems or the resulting contract», Magrani, 2019, p. 6.

[74] Por ejemplo, los algoritmos con capacidad de aprender desafían la responsabilidad del fabricante o del diseñador, máxime cuando hay un *machine learning*, Mittelstadt et al., 2016, 11.

[75] La complejidad de los sistemas de IA es mucho mayor que cualquier producto que hayamos tenido antes. La IA, el IoT y la robótica se caracterizan por combinar «la conectividad, la autonomía y la dependencia de datos para llevar a cabo tareas con poco o ningún control o supervisión humanos», Comisión Europea, 2020, *Informe sobre las repercusiones*, 2.

tase con procedimientos de certificación y auditoría que evaluasen la seguridad de su uso cuando van a adoptar decisiones autónomas, lo que estaría basado en los principios de seguridad, transparencia y explicabilidad.

4. **Reforzar la supervisión humana**. En los sistemas de IA con alta autonomía debería haber un deber de supervisión constante, de tal manera que los operadores pudieran monitorizar el comportamiento de la IA y detectar los riesgos o fallos para evitar que se produjeran daños. Este deber reforzado de supervisión humana tendría que ir acompañado de la posibilidad de intervención manual por parte de las personas que lo están monitorizando y todo ello supondría un deber de diligencia que serviría, en su caso, para apreciar la negligencia si fallara.
5. La **Responsabilidad compartida** también podría servir en estos casos para saber quiénes son los actores responsables cuando la IA actúa con mayor autonomía, lo que sería una exigencia de la transparencia y la explicabilidad.

3. 5. Daños colectivos y difusos

En los daños que puede causar la IA se pueden intensificar algunas categorías de daños, ya existentes y conocidos por el Derecho, pero que conviene tener en cuenta.

Por ejemplo, conviene recordar esquemáticamente el concepto de **intereses difusos** que estarían entre los intereses individuales y los colectivos. Interés individual es aquél que afecta y pertenece a un individuo de tal manera que su lesión puede dar lugar a RC. Por otro lado un interés colectivo se refiere a aquello cuya titularidad corresponde al colectivo y las medidas de reparación, si existen, serán exigibles por quien tenga representación del interés colectivo. Los intereses difusos se encuentran en un terreno intermedio por ser una lesión de algo individual que se causa a una colectividad. Pero una cosa son lesiones individuales hechas a un colectivo, lo que se puede solucionar a través de las acciones colectivas y otra aquellos casos en los que no llega a producirse una lesión individual que legitime una acción de RC, pero esa molestia se produce a un colectivo. Un sesgo algorítmico que genera una discriminación en sectores como el crédito, la sanidad o el empleo puede ocasionar una lesión individual a un colectivo de personas, cuando se les produce un daño concreto que pueden reclamar por RC y a la vez una lesión de intereses difusos de todas aquellas personas que *podrían* haber sufrido el daño, porque han sido discriminadas aunque no hayan llegado a recibir el daño.

En la discriminación a gran escala a través de decisiones algorítmicas o en las decisiones automatizadas que afectan a miles de personas se pueden producir daños a colectivos y lesiones de intereses difusos. En los supuestos de daños individuales que afectan a un colectivo de personas, la solución tendrá que ver con el régimen de **acciones colectivas** que ya existe, con mayor o menor acierto, en la UE.

En los supuestos de lesiones a intereses difusos se pueden estar poniendo en cuestión principios como los de Bienestar, Equidad, Diversidad e Inclusión.

Diversas actuaciones que se podrían tener en cuenta son:

1. **Establecimiento de criterios para cuantificar daños difusos.** Sería adecuado desarrollar criterios que permitieran cuantificar los daños difusos y graduales causados por la IA en contextos en los que el daño afectara a un grupo o colectividad. Estos criterios podrían basarse en la magnitud del impacto y en la repetición del daño, midiendo los efectos acumulativos de las decisiones de la IA. Tales criterios servirían para medir si se cumplen principios como: «*AIS must be designed and trained so as not to create, reinforce, or reproduce discrimination based on -among other things- social, sexual, ethnic, cultural, or religious differences*»[76].

2. **Auditoría algorítmica para identificar daños difusos.** Dentro de los sistemas de auditoria propuestos en otros apartados, podrían establecerse auditorías algorítmicas que pudieran identificar patrones de discriminación o perjuicio causados por la IA[77]. Además de los principios de bienestar y equidad, esto favorecería los de transparencia y explicabilidad[78].

3. En los **fondos de compensación** vistos en otros apartados, podría verse si podrían incluirse soluciones para los daños colectivos e incluso formas de compensación para las lesiones a los intereses difusos.

---

[76] IVADO, 2018, Principio 6 Equidad.

[77] Por ejemplo, Floridi et al. en su recomendación 4 proponen: «*Develop auditing mechanisms for AI systems to identify unwanted consequences, such as unfair bias, and (for instance, in cooperation with the insurance sector) a solidarity mechanism to deal with severe risks in AI-intensive sectors*», Floridi et al., 2018, p. 702.

[78] El principio de transparencia algorítmica «exige la visibilidad, la cognoscibilidad, la auditabilidad y la explicabilidad de los factores intervinientes en las decisiones tomadas con algoritmos a las personas que utilizan, regulan, y son afectadas por los sistemas de IA que emplean dichos algoritmos», Llano Alonso, 2024, p. 194.

3. 6. Daños futuros e inciertos

El principio de precaución subraya la necesidad de anticipar y evitar posibles riesgos futuros de la IA. Sin embargo, la incertidumbre inherente al desarrollo tecnológico hace difícil prever todos los efectos negativos, lo que complica la prevención y la responsabilidad por daños potenciales que aún no se han manifestado.

Es importante recordar que el concepto de daño es dinámico y ha variado a lo largo del tiempo, por lo que aplicamos, por ejemplo, nuestros Códigos Civiles a realidades que ni siquiera se podían haber soñado cuando se promulgaron los Códigos. Una de las grandes incertidumbres con las que tenemos que convivir es la necesidad de conocer los riesgos de la IA. Por ejemplo, no conocemos los riesgos para la salud mental de los usuarios que pueden generar las aplicaciones de IA que implican la colaboración con robots o sistemas de IA humanoides en el hogar o en entornos laborales[79]: es concebible que realidades que hoy no conocemos se consideren daños en el futuro. También puede haber daños importantes que no seamos capaces de prever[80].

Lo anterior está relacionado con principios como el de prevención del daño o seguridad. Algunas ideas que se podrían explorar son:

1. Establecimiento de **mecanismos de monitoreo y evaluación a largo plazo** para los sistemas de IA[81]. Esto implicaría una supervisión continua de los efectos de la IA, incluso años después de su implementación, para detectar daños potenciales antes de que se manifiesten completamente y evitar que las víctimas se vean afectadas sin protección.
2. **Revisión y actualización de las regulaciones**. Los tiempos de los legisladores deberían acomodarse a los tiempos de las realidades cambiantes. Baste recordar lo que ha costado en la UE la elaboración de la Ley de Inteligencia Artificial y cómo, durante su tramitación, surgieron novedosas aplicaciones como la IA generativa que no esta-

---

[79] Comisión Europea, 2020, *Report on the*, p. 9.

[80] Dentro de los asuntos a largo plazo el Principio 21 de Asilomar señala «*Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact*» Future of Life Institute, 2017.

[81] Por ejemplo las Recomendaciones de la UNESCO señalan: «Los Estados Miembros deberían alentar y promover la investigación colaborativa sobre los efectos de la interacción a largo plazo de las personas con los sistemas de IA, prestando especial atención a las consecuencias psicológicas y cognitivas que estos sistemas pueden tener en los niños y los jóvenes. Para ello deberían utilizarse múltiples normas, principios, protocolos, enfoques disciplinarios y un análisis de la modificación de las conductas y los hábitos, así como una cuidadosa evaluación de los impactos culturales y sociales posteriores», UNESCO, 2022, 129.

ba inicialmente contemplada en la Ley. Teniendo en cuenta los avances en estas materias, procedimientos legislativos que tardan años en culminar pueden nacer obsoletos[82]. Las propias Directrices señalan que la normativa, como método no técnico, debe someterse a evaluación constante[83].

Deberíamos tender hacia una regulación proactiva[84] y no simplemente reactiva con años de retraso.

Además, como señalaba el Principio 3 de Asilomar: «*3) Science-Policy Link: There should be constructive and healthy exchange between AI researchers and policy-makers*» (Future of Life Institute, 2017).

3. Diseño de **fondos de compensación** podrían estar diseñados para poder hacer frente a contingencias futuras que quizás todavía no se puedan prever.

4. BIBLIOGRAFÍA UTILIZADA

Abbott, R. B. (2018). The reasonable computer: Disrupting the paradigm of tort liability. *George Washington Law Review, 86* (1), 1-68.

Atienza Navarro, M. L. (2023). ¿Son necesarias reglas especiales para los daños causados por inteligencia artificial? En *Derecho e Inteligencia Artificial [El jurista ante los retos de la era digital]*. Thomson Reuters Aranzadi.

Benhamou, Y., y Ferland, J. (2021). Artificial intelligence & damages: Assessing liability and calculating the damages. En *Leading Legal Disruption: Artificial Intelligence and a Toolkit for Lawyers and the Law* (pp. 168-170). Thomson Reuters.

Borghetti, J.-S. (2019). Civil liability for artificial intelligence: What should its basis be? *La Revue des Juristes de Sciences Po,* (17), 1-30.

Bostrom, N., y Yudkowsky, E. (2014). The ethics of artificial intelligence. En K. Frankish & W. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 316-334). Cambridge University Press.

---

[82] Los diferentes tiempos del legislador y de los avances tecnológicos son un problema en materia de IA y de no fácil solución, porque la legislación necesita reflexión y debate. «*Even though the "legal lag" is more complex than it may seem, the speed of change, in particular, is still repeatedly a difficult challenge in relation to the inertia of traditional regulation. Legislative processes aimed at learning technologies with increasing agency require reflection, critical studies, and more knowledge in order to be able to find the desirable societal balances between various interests*», Larsson, 2020, p. 447.

[83] Grupo Independiente de Expertos, 2019, *Directrices éticas*, 103.

[84] «*Proactive measures and strict regulations are needed to ensure that AI development follows ethical guidelines and remains aligned with human values to prevent harmful outcomes*», (Bostrom y Yudkowsky, 2014)

Bradford, A. (2012). The Brussels effect. *Northwestern University Law Review, 107*(1), 1-68. https://scholarlycommons.law.northwestern.edu/nulr/vol107/iss1/1

Colmenarejo, A. B., Nannini, L., Rieger, A., Scott, K. M., Zhao, X., Patro, G. K., Kasneci, G., & Kinder-Kurlanda, K. (2022). Fairness in agreement with European values: An interdisciplinary perspective on AI regulation. En *AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 107-118). Association for Computing Machinery. https://doi.org/10.1145/3514094.3534158

Comisión Europea. (2020). *Libro blanco sobre la inteligencia artificial: Un enfoque europeo orientado a la excelencia y la confianza*. COM(2020) 65 final. https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52020DC0065

Comisión Europea. (2019). *Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones, Generar confianza en la inteligencia artificial centrada en el ser humano*. COM(2019) 168 final. https://eur-lex.europa.eu/legal-content/ES/TXT/DOC/?uri=CELEX:52019DC0168

Comisión Europea. (2020) Informe sobre las repercusiones en materia de seguridad y responsabilidad civil de la inteligencia artificial, el internet de las cosas y la robótica. COM/2020/64 final. https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:52020DC0064

Comisión Europea, *Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics*, Brussels, 19.2.2020 COM(2020) 64 final, p. 9. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0064

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*(1), 1-3. https://doi.org/10.1007/s10676-018-9450-z

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., y Vayena, E. (2018). AI4People: An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

García Teruel, R. M. (2021). El derecho de daños ante la inteligencia artificial y el machine learning: Una aproximación desde las recomendaciones del Parlamento Europeo y del Grupo de Expertos de la Comisión Europea. En F. Roca Guillamón (Ed.), *Cuestiones clásicas y actuales del derecho de daños*. Aranzadi.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Future Of Life Institute. (2017). Asilomar AI principles. https://futureoflife.org/ai-principles/

Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial. (2019). *Directrices éticas para una IA fiable*. Comisión Europea. https://doi.org/10.2759/177365

Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial. (2019). *Policy and investment recommendations for trustworthy AI.* Comisión Europea. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

IEEE. (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (1ª ed.). https://ethicsinaction.ieee.org/

Institut de Valorisation des Données (IVADO). (2018). *Montreal declaration for a responsible development of artificial intelligence.* https://www.montrealdeclaration-responsibleai.com/

IPSOS. (2023). *Global views on AI 2023.* IPSOS. https://www.ipsos.com/sites/default/files/ct/news/documents/2023-07/Ipsos%20Global%20AI%202023%20Report.pdf

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society, 7*(3), 437-451. https://doi.org/10.1017/als.2020.19

Llano Alonso, F. H. (2024). Ética(s) de la inteligencia artificial y derecho: Consideraciones a propósito de los límites y la contención del desarrollo tecnológico. *Derechos y Libertades, 51*, 177-199. https://doi.org/10.20318/dyl.2024.8587

Magrani, E. (2019). New perspectives on ethics and the laws of artificial intelligence. *Internet Policy Review, 8*(3). https://doi.org/10.14763/2019.3.1420

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., y Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1-21. https://doi.org/10.1177/2053951716679679

Müller, L. (2022). Domesticating artificial intelligence. *Moral Philosophy and Politics, 9*(2), 219-237. https://doi.org/10.1515/mopp-2020-0054

OECD. (2019). *Recommendation of the Council on Artificial Intelligence.* OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OdiseIA, PwC, Google, Microsoft, IBM, y Telefónica. (2022). *Guía de buenas prácticas en el uso de la inteligencia artificial ética.* Recuperado de https://www.digitalidades.org/guia-de-buenas-practicas-en-el-uso-de-la-inteligencia-artificial-etica/

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown Publishing Group.

Oxford Internet Institute. (2020). *Global attitudes towards AI, machine learning & automated decision making.* University of Oxford. https://www.oii.ox.ac.uk/research/global-attitudes-ai-2020.pdf

Parlamento Europeo. (2022). *Resolución del Parlamento Europeo, de 3 de mayo de 2022, sobre la inteligencia artificial en la era digital (2020/2266(INI)).*

Parlamento Europeo. (2017). *Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de derecho civil sobre ro-*

bótica (2015/2103(INL)). https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html

Parlamento Europeo. (2020). *Informe con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial* (2020/2014(INL)).

Parlamento Europeo, Resolución del Parlamento Europeo, de 20 de octubre de 2020, con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial (2020/2014(INL)) https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_ES.html#title2

Propuesta de Directiva del Parlamento Europeo y del Consejo relativa a la adaptación de las normas de responsabilidad civil extracontractual a la inteligencia artificial (Directiva sobre responsabilidad en materia de IA). COM/2022/496 final https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A52022PC0496

Propuesta de Directiva del Parlamento Europeo y del Consejo sobre responsabilidad por los daños causados por productos defectuosos. COM/2022/495 final https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A52022PC0495

Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.° 300/2008, (UE) n.° 167/2013, (UE) n.° 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial) ELI: http://data.europa.eu/eli/reg/2024/1689/oj

Scherer, M. U. (2016). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology*, *29*(2), 353-400. http://dx.doi.org/10.2139/ssrn.2609777

Selbst, A. D. (2020). Negligence and AI's human users. *Boston University Law Review*, *100*, 1315-1376.

Stamboliev, E., y Christiaens, T. (2024). How empty is trustworthy AI? A discourse analysis of the ethics guidelines of trustworthy AI. *Critical Policy Studies*. https://doi.org/10.1080/19460171.2024.2315431

Susskind, R. (2019). *Online courts and the future of justice*. Oxford University Press.

UNESCO. (2022). *Recomendación sobre la ética de la inteligencia artificial*. https://unesdoc.unesco.org/ark:/48223/pf0000380455_spa

Van Wynsberghe, A. (2016). *Healthcare robots: Ethics, design and implementation*. Routledge.

Velasco Perdigones, J. C. (2021). Fundamentos para la atribución de responsabilidad civil extracontractual en la "era tecnológica". En J. Cruz Ángeles, M. Novo Foncubierta, B. Martín Novo, y M. Paniagua Zurera (Coords.), *El sistema jurídico ante la digitalización. Estudios de derecho privado*. Tirant lo Blanch.

Vladeck, D. C. (2014). Machines without principals: Liability rules and artificial intelligence. *Washington Law Review*, *89*(1), 117-150. https://digitalcommons.law.uw.edu/wlr/vol89/iss1/6

Wagner, G. (2018). Robot liability. *SSRN*. http://dx.doi.org/10.2139/ssrn.3198764

La inteligencia artificial tiene el potencial de transformar productos, servicios y procedimientos en multitud de sectores económicos y en relación con muchos ámbitos de la sociedad. Sin embargo, también puede generar un sinfín de riesgos que, de producir daños, habrán de ser reparados. La Unión Europea no ha sido ajena a estos riesgos, y por ello ha pretendido y sigue pretendiendo crear un marco jurídico protector. Dentro de este contexto, se sitúa la aprobación del Reglamento (UE) 1689 del Parlamento y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial -RIA-, como sendas Propuestas de Directiva, de inminente aprobación, sobre responsabilidad civil de productos defectuosos y sobre responsabilidad civil por daños causados por la inteligencia artificial. Partiendo de tales postulados, en la presente obra se han seleccionado aquellos sectores donde, por su mayor proyección, novedad o complejidad, merece ser analizada la interrelación entre la tecnología de la inteligencia artificial y el Derecho de daños. Para ello, se ha podido contar con un elenco de especialistas en el sector, que sin duda hace de la obra resultante una aportación doctrinal de indudable utilidad.

Con carácter particular, entre los sectores seleccionados, destaca por su trascendencia, el de la salud digital, donde problemáticas relacionadas con sistemas inteligentes para la prevención de enfermedades, ya sea a iniciativa del profesional de la medicina, o al margen de él -uso de werables y servicios digitales-, o por infracciones de los datos personales de salud, pueden determinar, si bien a través de distintos cauces normativos, posibles vías de reclamación indemnizatoria.

En el campo quirúrgico, la "cirugía 4.0", que integra la cirugía robótica y personalizada, por su creciente implantación, ha merecido una especial consideración en la obra.

Se efectúan igualmente amplías consideraciones acerca de la transparencia y explicabilidad para prevenir la discriminación algorítmica en el uso de los sistemas de inteligencia artificial.

Dentro de los sectores con mayor implementación de las tecnologías de inteligencia ha sido objeto de consideración así mismo el uso de vehículos autónomos, incluida su problemática en la vertiente del Derecho internacional privado.

Situados en el marco normativo que proporciona el Reglamento de Inteligencia artificial -RIA- se efectúan correspondientes análisis acerca de la categorización del riesgo que el mismo contempla, y donde se observa un régimen jurídico tendente a salvaguardar los riesgos más graves por el empleo de los sistemas de inteligencia artificial; en particular, en la salud, seguridad y derechos consagrados en la Carta Europea de Derechos Fundamentales. De igual forma las implicaciones jurídicas que despliega la inteligencia artificial generativa por infracciones normativas del Derecho de protección de datos personales. Se incluyen también los rasgos que deben estar presentes en el seguro de responsabilidad civil profesional de los operadores de inteligencia artificial, a partir de las previsiones normativas del referido Reglamento de Inteligencia Artificial.

9 788410 707085