ALLENDE UNA TEORÍA GENERAL DE LAS GARANTÍAS JURÍDICAS PARA UNA INTELIGENCIA ARTIFICIAL CONFIABLE[1]

Pere Simón Castellano*

1 El presente trabajo de investigación se enmarca como resultado del proyecto con referencia TED2021-129356B-I00, intitulado "Sobre las bases normativas y el impacto real de la utilización de algoritmos predictivos en los ámbitos judicial y penitenciario", Ministerio de Ciencia e Innovación, convocatoria 2021 - Proyectos de Transición Ecológica y Transición Digital y en el marco Proyecto Retos MICINN (PID2022-136439OB-I00).

* Profesor Titular de Derecho Constitucional de la Universidad Internacional de la Rioja (UNIR).

I. A MODO DE INTRODUCCIÓN: DE LA INSUFICIENCIA DE LAS NORMAS SECTORIALES Y LA NECESIDAD DE UN ENFOQUE DE RIESGO *AD HOC*

El desarrollo de tecnologías disruptivas de alto valor añadido y, más concretamente, de aquellas soluciones o herramientas que son susceptibles de ser clasificadas como Inteligencia Artificial[2] (en adelante, IA), se plantea como solución en el debate público y en términos tanto de necesidad como de oportunidad. No es de extrañar que, en España, siguiendo la línea marcada por Canadá[3] en 2017, la Estrategia Nacional de Inteligencia Artificial actúe como uno de los ejes principales de la Agenda España Digital 2026, que tiene como principal objetivo crear un marco de referencia estable para el desarrollo de una IA inclusiva, sostenible y confiable.

El mercado y amplios sectores de la actividad pública y privada han puesto su foco de atención sobre la brillante coyuntura que supone, en términos de automatización y digitalización, abrazar la verosimilitud de que las máquinas u algoritmos puedan emular o bien el proceso cognitivo humano o bien el proceso decisorio derivado; unas acciones o decisiones automatizadas que a vueltas pueden proyectar sus efectos incidiendo en muchos aspectos de la vida cotidiana de los *cives*, de las que pueden derivar incluso consecuencias jurídicas.

Como reacción, y también como derivada del progresivo afán por parte del regulador, especialmente el europeo –en la carrera para llegar primero a nivel internacional, como ya se hizo con cierto éxito con la normativa de protección

2 Una definición amplia de IA incluye cualquier tipo de desarrollo tecnológico con capacidad para emular procesos cognitivos y decisorios humanos, y precisamente por ello, no necesariamente racionales.

3 Desde 2017, cuando el Canadá se convirtió en el primer país a adoptar una estrategia nacional de inteligencia artificial, al menos 60 países han adoptado políticas (declaraciones, marcos, directrices de la industria o principios) centradas en la IA, para trabajar junto con el sector tecnológico, la academia y la sociedad civil para un diseño y aplicación de la IA responsable, de confianza y ético. En este sentido, en Cataluña se puso en marcha en febrero de 2020 la Estrategia de Inteligencia Artificial de Cataluña, para fortalecer el ecosistema catalán en IA teniendo muy presente un eje sobre "Ética y Sociedad".

de datos[4]– para delimitar el empleo de la tecnología más disruptiva, pueden apreciarse dos tendencias que confluyen. De un lado, la creciente labor por parte de académicos que publican artículos sobre IA que, por lo general, se limitan a reproducir cuestiones terminológicas o semánticas, alertar o abrazar el avance tecnológico –al más puro estilo apocalípticos e integrados, siguiendo aquí a la categorización que formuló Eco (2004)– y aportar relativamente poco al debate regulatorio o a subsectores específicos que requieren de estudio y análisis tanto a nivel teórico como práctico. Paga la pena entonces hacer referencia a las pocas excepciones de la premisa anterior[5]. Por otro lado, destaca la proliferación de guías de principios éticos (Jobin, 2019), o lo que es lo mismo, directrices y consideraciones éticas como declaraciones de principios que a menudo resultan demasiado complejas y difíciles de entender puesto que están escritas para usuarios académicos o técnicos. La doctrina ha tenido la ocasión de criticar este extremo, al dificultar su implementación por parte de diferentes operadores interesados tales como las empresas, la Administración, los organismos profesionales y las organizaciones de la sociedad civil, retrasando su adopción y aplicación como un elemento de valor de innovación responsable (Morley, 2019).

La confluencia de esas dos tendencias –proliferación de estudios doctrinales y de cartas, directrices y guías de principios éticos y jurídicos de la IA– se agrava en la medida que en la mayoría de las ocasiones estas son resultado de procesos en los que no se han incorporado o participado potenciales personas usuarias, destinatarias o afectadas por los propios sistemas de IA.

En este escenario, nada halagüeño de entrada, un servidor ha dedicado sus esfuerzos en defender que (1) el debate sobre la IA debe centrarse en determinar qué IA y para qué usos, basando su aplicación en un modelo de responsabilidad proactiva, con medidas por defecto y en el diseño, con modelos de prevención y certificación que ya funcionan en otros ámbitos que desde hace décadas operan bajo estándares internacionales de normalización

4 Nos referimos al Reglamento (UE) 2016/679, del Parlamento Europeo y del Consejo, de 27 de abril de 2016, General de Protección de Datos (en adelante, RGPD).

5 Resultan referencia obligada los trabajos de Cerrillo (2020), Cotino (2021, 2022a y 2022b), Nieva (2018), Roig (2020), Presno (2022), Rallo (2020), Medina (2022), Mantelero (2018 y 2020), Lipton (2018), Kroll (2021), Madaio (2020), Simón (2021, 2022a y 2022b), Marchena (2022) y Miró (2018 y 2020). Esta nota no puede ser exhaustiva si tenemos en cuenta la ingente literatura publicada sobre IA en los últimos años.

(Simón, 2021); (2) debemos avanzar con la creación de un marco normativo europeo de referencia basado en un enfoque de riesgo, con medidas de *hard law* –establecimiento de líneas rojas– y una definición de medidas preventivas con distinta intensidad por lo que se refiere a las obligaciones y salvaguardas en función de los niveles de riesgo concretos de la tecnología –naturaleza– a aplicar en cuestión y de su contexto y alcance (Simón, 2022a). El presente capítulo parte, así, de la constatación, tal y como ya se ha argumentado en otros foros (Simón, 2022b), de la insuficiencia del marco normativo en protección de datos y transparencia del sector público para ofrecer una respuesta coherente y sistemática a la problemática que ofrece la aplicación e implementación de sistemas basados en IA. Y pretende ir más allá sosteniendo la necesidad de que esa respuesta esté vinculada inextricablemente a la dignidad humana, al libre desarrollo de la personalidad y a la dimensión objetiva de los derechos (Cotino, 2022). Como derivada de lo anterior, resulta más necesario que nunca formular una taxonomía concreta de las garantías frente al uso de los sistemas de IA, tratando de delimitar con claridad conceptos como transparencia y explicabilidad, a menudo confundidas especialmente por parte de juristas, separando claramente sus objetos y propiedades, ejercicio que en cualquier caso debe leerse de forma crítica y comprendiendo que, para cada desarrollo o implementación tecnológica, deberá analizarse el nivel de riesgo residual resultante en base a las garantías que aplican, descartando aquellas que resulten ineficaces, inidóneas o simplemente imposibles[6].

En virtud de lo anterior, planteamos como hipótesis que se pretenden contrastar o refutar a través del presente trabajo, las siguientes:

Primera y punto de partida.- El desarrollo y empleo de sistemas de IA requiere de una normativa específica basada en un enfoque de riesgo, fruto de la insuficiencia de regulaciones sectoriales.

Segunda.- Las garantías frente al uso de sistemas de inteligencia artificial deben clasificarse y sistematizarse mediante un esfuerzo de dogmática aplicada; es necesario conceptuar y sistematizar los factores y (sub)propiedades de garantías tales como la seguridad, transparencia, explicabilidad, etc. aunque sólo sea por mero pragmatismo y razones utilitarias.

6 Planteamiento que ha sido objeto de estudio y foco principal del trabajo (Simón, 2022, en prensa).

Tercera.- Debe superarse la confusión conceptual entre transparencia y explicabilidad, siendo garantías con autonomía conceptual y propiedades diferenciadas. La primera exige obligaciones de publicidad activa y genéricas sobre la herramienta tecnológica, mientras que la segunda se refiere a la capacidad de la propia tecnología de explicar las razones o motivación de su decisión en el caso concreto. Ambas son garantías que persiguen un fin mayor, más allá de la propia transparencia o explicabilidad, que no es otro que el derecho de defensa, a recurrir o a obtener una segunda oportunidad frente al algoritmo. Esto último, el derecho de defensa o al recurso, también llamado por la Comisión Europea como el derecho a una segunda oportunidad, es la verdadera línea roja de *hard law* que subyace, esto es, el fin u objetivo principal que el regular pretende garantizar.

Cuarta.- Otras garantías como la seguridad –en especial, la trazabilidad como subpropiedad–, la participación humana –garantías subjetivas– o la supervisión *ex ante* y *ex post* –garantías institucionales– juegan un papel clave si de lo que se trata es de conseguir una IA confiable, responsable y ética.

II. QUÉ GARANTIZAR: ¿UN FUNCIONAMIENTO TRANSPARENTE O LA POSIBILIDAD DE RECURRIR Y DEFENDERSE FRENTE A UNA DECISIÓN ALGORÍTMICA?

Tradicionalmente, la respuesta de los operadores jurídicos y su principal queja frente al empleo en la práctica de los sistemas de inteligencia artificial se ha vehiculado y centrado en una suerte de derecho a conocer el código algorítmico y, en el mejor de los casos, también su funcionamiento[7].

Sin embargo, la transparencia o el acceso íntegro al código fuente *stricto sensu* no es un fin en sí mismo y no actúa a modo de panacea frente a los posibles usos no deseados de los algoritmos. Lo expresa claramente Medina (2022: 169) cuando advierte que «el objetivo de lograr la rendición de cuentas de las decisiones algorítmicas imponiendo su explicación a los afectados no es sino

7 Especialmente por parte de la doctrina administrativista. Véanse los trabajos de Cerrillo (2020), Boix (2020a); para un enfoque vinculado a la protección de datos y conectado con el derecho de acceso del art. 15 del RGPD, véase Medina (2022: 162 y ss.).

una solución ingenua, el vano empeño de alcanzar una elusiva *fata morgana*, y, por lo tanto, que con la pretensión de hallar en el RGPD un amplio derecho a la explicación se corre el riesgo de desembocar en una suerte de falacia de la transparencia».

De lo que se trata no es de acceder al código algorítmico o de que su diseño y lógica sea necesariamente transparente[8], sino que las personas no sean objeto de una decisión automatizada, no humana, de la que se deriven efectos jurídicos sin que esta sea capaz de defenderse o poder reaccionar con garantías frente a ella. Esa es la preocupación del legislador europeo, que ya de forma temprana y en un ámbito más limitado –normativa de protección de datos–, introdujo un literal como el del art. 22.3 del RGPD, que se refiere al «derecho a obtener intervención humana por parte del responsable, a expresar su punto de vista y a impugnar la decisión».

Como se observa, saber la lógica y garantizar el acceso al código de los sistemas de inteligencia artificial puede ayudar, en ocasiones, a ese objetivo integral y final de garantizar que nadie será objeto de una decisión automatizada –cuyos responsables sean empresas privadas o el sector público[9]– sin posibilidad de

8 Tampoco vehicular la transparencia a través del derecho de acceso individual previsto *ex* art. 15 del RGPD. De nuevo, al igual que hemos comentado anteriormente de forma crítica en relación con la protección de datos en general, se trata de un enfoque limitado e insuficiente, y el legislador consciente de estas dificultades ha tratado de armonizar en una única normativa o instrumento de aplicabilidad directa una respuesta *ad hoc*. Véase la Propuesta de Reglamento del Parlamento Europeo y del Consejo, de fecha 21 de abril de 2021.

9 Este enfoque también permite incorporar a la ecuación, en relación con el eventual empleo de los algoritmos en el sector público, el principio y derecho a la buena administración que, derivado de los arts. 9.3 y 103 de la Constitución, pero más concretamente del art. 41 de la Carta de los Derechos Fundamentales de la Unión Europea, dice la Sala Tercera del Tribunal Supremo, «ha adquirido el rango de derecho fundamental en el ámbito de la Unión, calificándose por algún sector doctrinal como uno de los derechos fundamentales de nueva generación» que «es algo más que un derecho fundamental de los ciudadanos, siendo ello lo más relevante; porque su efectividad comporta una indudable carga obligación para los órganos administrativos a los que se les impone la necesidad de someterse a las más exquisitas exigencias legales en sus decisiones, también en las de procedimiento». Véanse, respectivamente, el ATS (Sala 3a) núm. 3593/2022, de 16 de marzo de 2022, ECLI:ES:TS:2022:3593A, y la STS (Sala 3a) núm. 1667/2020, de 3 de diciembre de 2020, ECLI:ES:TS:2020:4161. Difícilmente puede hablarse de buena administración si esta adopta decisiones basadas en algoritmos cuya lógica se desconoce o no permitiendo que los ciudadanos puedan formular recursos y defenderse con ciertas garantías frente a las proyecciones o propuestas que ofrecen las herramientas tecnológicas.

recurrir con garantías. Sin embargo, en función de la tecnología aplicada y sus usos, el acceso al código puede resultar inidóneo, lo que exige, de un lado, reinterpretar la transparencia hacía conceptos que la atomizan como los relativos a la legibilidad, la trazabilidad, la testabilidad, la auditabilidad o la verificabilidad, y del otro apoyarse en otras garantías específicas propias de la seguridad de los sistemas de gestión de la seguridad de la información y de la metodología de los análisis de riesgos propios de los sistemas de cumplimiento normativo[10].

Se trata de un enfoque superior, más grande y general, con una visión sistemática de la normativa y principios aplicables a las soluciones basadas en algoritmos de inteligencia artificial, que nos muestra que la transparencia es sólo un extremo más, una garantía de garantías, como se verá, necesariamente interrelacionadas con nuevos paradigmas como la seguridad y la explicabilidad o el «derecho a la segunda oportunidad», tal y como lo ha bautizado el Parlamento Europeo[11].

---

10 En esta misma dirección resulta de sumo interés traer a colación el *Algorithmic Transparency Standard* aprobado en Reino Unido por el *UK Central Digital and Data Office*, publicado el 29 de noviembre de 2021, disponible en Internet: https://www.gov.uk/government/collections/algorithmic-transparency-standard (última consulta el 9 de julio de 2022). Se trata de un estándar que no es preceptivo salvo para el sector público en relación con el empleo de herramientas basadas en algoritmos para la toma de decisiones. Un estándar de «trasparencia» aunque las medidas que incluye van mucho más allá: en primer lugar (1) los organismos públicos deben explicar el funcionamiento y el problema que tratan de resolver empleando herramientas basadas en inteligencia artificial para, a continuación cumplir con un completo (2) listado de requisitos y detalles técnicos, lo que incluye la definición del propietario del proceso, el alcance de la herramienta –diseñada para qué fines, el mantenimiento previsto, la arquitectura del sistema, tipo de modelo, uso recurrente en el tiempo–, la explicación relativa al cómo la herramienta afecta a la toma de decisiones y la participación humana vía auditorías y procesos de revisión y mejora continua, el cumplimiento de las obligaciones en materia de protección de datos –descripción de categorías, bases de legitimación, el ciclo de vida del tratamiento, la definición de cómo han sido recabados los datos, los acuerdos con corresponsables y encargados en el caso que les hubiere, quién tiene acceso a los datos y plazos de conservación–, los informes de las evaluaciones de impacto –ético, discriminación, proporcionalidad y protección de datos–, la descripción de riesgos concretos –discriminación, sesgos y daños o perjuicios– y los controles previstos y en marcha para mitigar los riesgos identificados, con niveles de riesgo residual dentro de un umbral aceptable.

11 Sobre el posible «uso malintencionado de la inteligencia artificial y los derechos fundamentales», el Parlamento Europeo «insta a la Comisión a que tome nota de los retos sociales derivados de las prácticas resultantes de la clasificación de los ciudadanos; subraya que los ciudadanos no deben ser objeto de discriminación en función de su clasificación y que deben tener derecho a una segunda oportunidad». Véase la Resolución del Parlamento Europeo, de 12 de febrero de

III. EL ENFOQUE DE RIESGO PROPUESTO POR LA UE

El objetivo de la propuesta europea de Reglamento sobre IA[12] cuya creación se propone es garantizar que los ciudadanos europeos puedan confiar en lo que la inteligencia artificial puede ofrecer**,** permitiendo las mejoras técnicas a la par que reforzando la tutela de los derechos fundamentales frente a los riesgos que puede comportar el uso de herramientas o sistemas basados en las tecnologías más vanguardistas. Se persigue acabar con ese marco jurídico atomizado que ha sido analizado anteriormente, combinando y armonizando las distintas disposiciones normativas con una vocación integradora y coherente, junto a un nuevo plan coordinado que incorpora la participación de los distintos Estados miembros. Con ello se pretende, de un lado, garantizar la seguridad y los derechos fundamentales de las personas y las empresas, y del otro, reforzar la adopción e impulsar la inversión y la innovación en materia de inteligencia artificial en toda la Unión Europea.

La normativa en fase de tramitación parte de un enfoque horizontal con una definición amplia de sistema de inteligencia artificial, lo que permite garantizar la cobertura de las reglas establecidas a cualquier avance cercano a la inteligencia artificial, es decir, incluyendo los más básicos sistemas de procesamiento y lectura del lenguaje o imágenes, y también los sistemas expertos. El fin es asegurar la neutralidad tecnológica, lo que exige una definición amplia, flexible y dinámica de inteligencia artificial. La protección se amplía así a cualquier algoritmo, técnica, máquina o androide que replique una acción o un objetivo humanos, generando resultados como contenidos, predicciones, recomendaciones o decisiones que influyen en los entornos con los que esta interactúa.

La propuesta no distingue grandes empresas de pequeñas y medianas empresas (PYMES), por lo que se aplicará igualmente a todas. Más concretamente, se aplicará (1) a todos los proveedores que operen en la Unión Europea,

---

2019, sobre una política industrial global europea en materia de inteligencia artificial y robótica (2018/2088(INI)), disponible en Internet: https://www.europarl.europa.eu/doceo/document/TA-8-2019-0081_ES.html (última consulta el 9 de julio de 2022).

12 Nos referimos a la Propuesta de Reglamento del Parlamento Europeo y del Consejo, de fecha 21 de abril de 2021, citada anteriormente.

independiente si están o no establecidos en la Unión; (2) a todos los usuarios localizados en la Unión Europea, esto es, a cualquier persona –natural o jurídica–, autoridad gubernamental, agencia o entidad que utilice un sistema de inteligencia artificial en sus actividades principales o profesionales –se excluye utilización personal y no profesional–; (3) a los proveedores y usuarios localizados en otros países cuando el producto del sistema, es decir, el *output* tecnológico sea utilizado en cualquier país de la Unión Europea.

No será de aplicación, en cambio, a la inteligencia artificial desarrollada o utilizada exclusivamente para propósitos militares. Tampoco tendrá efectos para las autoridades gubernamentales de países de fuera de la Unión Europea, cuando la utilización de los sistemas de inteligencia artificial es amparada por acuerdos internacionales para cooperación judicial y entre fuerzas policiales.

La propuesta, además, ni altera la responsabilidad jurídica de los proveedores de servicios de intermediación, que será regulada por la *Digital Services Act*[13], ni regula responsabilidad jurídica general. Esa concepción y visión amplia, con un enfoque horizontal, también proyecta sus efectos sobre los obligados al cumplimiento. Los sectores a los que se aplicará no están definidos *ex lege*, no se formula un *numerus clausus* o lista cerrada y, en definitiva, su aplicación será obligatoria para todos aquellos sectores en los que se utiliza o se puedan utilizar los sistemas de inteligencia artificial, desplazando o condicionando, cuanto menos, la normativa sectorial que, por lo general, es incipiente y deberá adaptarse a las reglas generales establecidas en la propuesta europea.

13 Normativa de suma importancia por lo que se refiere a la responsabilidad de los prestadores, adoptando un modelo de corregulación, que encaja con los actuales esfuerzos voluntarios de las redes y plataformas. Se trata de una norma de *hard law* que respalda el poder y las propias acciones de las plataformas, estimulando la adopción de códigos de conducta –art. 35– y regulando los instrumentos, herramientas y órganos de garantía. Además, para las grandes plataformas tecnológicas y prestadores se introducen obligaciones relativas a una acción proactiva o *ex ante* de «evaluación de riesgos sistémicos» –arts. 25 y ss.–, algunas de ellas destinadas a evitar ciertas fórmulas de desinformación. En función de los resultados de la evaluación y de los niveles de riesgo residual, las plataformas deberán adoptar medidas de reducción de riesgo a futuro –art. 27– o realizar auditorías independientes –art. 28–, garantizando en cualquier caso vías de impugnación y tutela de los derechos de los usuarios frente a las decisiones de la plataforma. Véase la propuesta de Reglamento de Servicios Digitales (en adelante, empleando las siglas de su acrónimo en inglés, DSA), y la Resolución del Parlamento Europeo de 5 de julio de 2022, disponible en Internet: https://www.europarl.europa.eu/doceo/document/TA-9-2022-0269_ES.html (última consulta el 9 de julio de 2022). Véanse también las contribuciones de Cotino, L. (2022b: 199-238); Barata, J. (2021).

El futuro Reglamento incorpora en su Título V ciertas medidas destinadas a fomentar la innovación. Así, las autoridades competentes de los Estados miembros o el Supervisor Europeo de Protección de Datos podrán establecer *regulatory sandboxes* para estimular el desarrollo, prueba y validación de sistemas de inteligencia artificial innovadores por un período limitado, antes de su implementación o comercialización[14]. Los *sandboxes* no afectarán los poderes de supervisión y corrección de las autoridades competentes, y tampoco al régimen de responsabilidad jurídica por daños a terceros que resulten de los experimentos.

Los Títulos VI, VII y VIII, por su parte, se encuentran dedicados a los mecanismos de gobernanza y de aplicación del Reglamento. Respecto de la gobernanza tiene especial relevancia la creación del Comité Europeo de Inteligencia Artificial, compuesto por representantes de los Estados miembros, de la Comisión Europea y el Supervisor Europeo de Protección de Datos, sobre el que volveremos con mayor detalle en el epígrafe IV.5 del presente trabajo. Los sistemas de cumplimiento normativo y gobernanza previstos en la propuesta normativa incluyen un sistema de notificación a la autoridad nacional competente para efectos de conformidad, la necesidad de que los Estados miembros establezcan una autoridad nacional de certificación, la creación de sistemas de evaluación, certificación y registro –base de datos– de los sistemas de alto riesgo, sistema de monitorización del mercado, sistema de notificación de incidentes, designación de autoridades competentes de control de ámbito nacional, entre otros.

Por lo que se refiere al régimen sancionador, la propuesta establece que los Estados miembros deberán establecer las reglas relativas a las sanciones aplicables, incluidas las multas administrativas, por incumplimiento del Reglamento, y que estas deben ser en todo caso eficaces, proporcionales y disuasivas. La propuesta, siendo consecuente con las competencias europeas, establece límites a las sanciones, que en caso de ser consecuencia de un incumplimiento de las prohibiciones totales o de las obligaciones relativas a bases de datos, pueden alcanzar hasta los treinta millones de euros o hasta el 6% de la facturación anual global.

14 Véase Ranchordas (2021).

La doctrina nacional (Cotino *et al.*, 2021) e internacional (Smuha *et al.*, 2021) ha criticado parcialmente la propuesta, formulando enmiendas y sugerencias de mejora, con el fin de iluminar al legislador europeo, poniendo de relieve posibles lagunas relativas a la protección de los derechos de los ciudadanos frente a los daños causados por la tecnología, especialmente por la falta de reconocimiento de un derecho a obtener indemnización.

La Asociación de Consumidores Europeos (por sus siglas en inglés y en adelante, BEUC) denunció el mismo día de la presentación de la propuesta de Reglamento la debilidad con la que dicen esta protege los derechos de los consumidores, al ser demasiado dependiente de las propias valoraciones de la industria y contemplar demasiadas excepciones[15].

A esta valoración proveniente de las asociaciones, se ha sumado una crítica de corte más académico y profesional, que detalla mejor las ausencias y olvidos de la propuesta del legislador europeo. El grupo de expertos OdiseIA indica que la música de la propuesta presentada por la Comisión Europea suena bien, pero falta concretar más, especialmente desde el punto de vista de la ciudadanía, pues no se prevén mecanismos especiales de garantía ante las nuevas instituciones o ante los órganos judiciales por parte de los afectados. Así, en una publicación reciente de OdiseIA (Cotino *et al.*, 2021), se señala que el enfoque basado en niveles de riesgo es adecuado y coherente, pero sería necesario también añadir una regulación de usos específicos en aras de aumentar su precisión y eficiencia; así como reflexionar sobre el modelo de exigencia de cumplimiento que inspira el Reglamento, en la medida en que variar el nivel de exigencia según el tamaño de la empresa puede tener como consecuencia situaciones indeseadas, como que grandes empresas subcontraten los servicios de otras más pequeñas y esquivar así un control más exhaustivo.

La principal conclusión que ofrece el informe del grupo OdiseIA es que en la tramitación de la propuesta habría que ajustar y definir mejor los usos de inteligencia artificial prohibidos y en su caso prever un mecanismo de actualización, una acción que en paralelo debería venir acompañada de una mejor

15 Véanse las quejas y crítica en la propia web de la BEUC, disponible en Internet: https://www.beuc.eu/publications/eu-proposal-artificial-intelligence-law-weak-consumer-protection/html (última consulta el 9 de julio de 2022).

delimitación de los sistemas de alto riesgo y, también, de un mecanismo de actualización *ad hoc* (Cotino *et al.*, 2021).

Por su parte, otros autores han puesto el foco de mejora en el hecho que la propuesta no incluye disposición alguna sobre el derecho a indemnización en situaciones en las que el sistema de inteligencia artificial infrinja lo previsto en ella (Miguel, 2021). Una opción que se atribuye a la decisión de la Comisión de revisar la Directiva sobre responsabilidad por los daños causados por productos defectuosos para adaptarla a las exigencias de las nuevas tecnologías, incluida la Inteligencia Artificial.

Finalmente, resulta necesario traer a colación la rápida respuesta del grupo de investigación del *Leads Lab* de la Universidad de Birmingham, en el Reino Unido (Smuha *et al.*, 2021). Los autores aplauden una parte importante de la propuesta, destacan el compromiso de hacer frente a los riesgos de la inteligencia artificial estableciendo un conjunto de obligaciones y un organismo público para su control; celebran una clasificación de sistemas de inteligencia artificial basada en niveles de riesgo más refinada que la fórmula optada en Libro Blanco de la Comisión sobre Inteligencia Artificial[16]; valoran muy positivamente la introducción de prácticas de inteligencia artificial prohibidas o la creación de una base de datos europea para los sistemas de alto riesgo.

Sin embargo, los citados autores consideran también que, tal y como está redactada, la propuesta no proporciona una protección adecuada de los derechos fundamentales, ni tampoco una protección suficiente para mantener el Estado de Derecho y la democracia, y por lo tanto no garantiza una inteligencia artificial basada en los principios de legalidad, ética y robustez (Smuha *et al.*, 2021). Más concretamente, establecen recomendaciones para los diferentes títulos de la propuesta, en relación con tres ejes dinámicos: la necesidad de garantizar la prevalencia de los derechos fundamentales de los ciudadanos en caso de conflicto; la tutela inexistente de los derechos fundamentales; por último, la falta de garantías sobre la transparencia y la rendición de cuentas de los sistemas de inteligencia artificial. También proponen añadir en la versión final del Reglamento un derecho explícito de reparación para las personas que

16 Véase el Libro Blanco sobre la inteligencia artificial – un enfoque europeo orientado a la excelencia y la confianza, adoptado en Bruselas, 19.2.2020. COM (2020) 65 final.

se vean sometidas a sistemas de inteligencia artificial no conformes, similar al derecho de indemnización de los titulares de los datos personales –aunque como ya hemos indicado *ut supra* este último ha tenido un desarrollo limitado y una aplicación práctica más que reducida–.

Con todo, la propuesta se limita a dar una respuesta proporcional al riesgo generado por los sistemas de inteligencia artificial, de modo que se limita a regular aquellos extremos en los que se generan elevados riesgos. El enfoque de riesgo condiciona la estructura del reglamento que se propone, en base a la clasificación de riesgos en cuatro niveles: inadmisible, alto, limitado y mínimo.

El artículo 5 de la propuesta de Reglamento objeto de estudio está dedicado a la prohibición de sistemas de inteligencia artificial cuando se considera que generan riesgos inadmisibles por contravenir los valores de la Unión, en particular, al facilitar la vulneración de derechos fundamentales. La propuesta se refiere a sistemas de inteligencia artificial que permitan identificar biométricamente –huella dactilar, reconocimiento facial, etc.– de forma remota en espacios públicos con fines policiales; a instrumentos que consigan perfilar y puntuar socialmente a las personas por parte de las autoridades públicas; al empleo de técnicas subliminales que pueden conducir a la manipulación de personas generando el riesgo de causar daños físicos o psicológicos a la persona en cuestión o terceros; los sistemas que pretenden aprovecharse de la especial vulnerabilidad de determinados grupos de personas. La prohibición por riesgo inadmisible no se refiere al diseño tecnológico, sino a la introducción en el mercado, la puesta en servicio y el uso de los sistemas de inteligencia artificial en cuestión en el conjunto de la Unión.

Los sistemas de riesgo alto, por su parte, son objeto de regulación en el Título III de la Propuesta, que abarca los artículos 6 a 51 y sus correspondientes anexos. La calificación como de alto riesgo se lleva a cabo en función de su potencial lesivo en la seguridad de las personas o respecto de sus derechos fundamentales. Se trata de sistemas cuya utilización es permitida, pero sujeta a un alto grado de control por las autoridades en virtud de los riesgos potenciales a la salud, seguridad y a otros derechos. Algunos ejemplos de inteligencia artificial que se consideran de riesgo alto son los siguientes: máquinas o robots operados autónomamente con sistemas de inteligencia artificial, incluidos los juguetes; dispositivos médicos; aviación civil; vehículos

automotores; identificación biométrica; operación y gestión de infraestructuras criticas; aplicación en el sector de la educación; empleo y recursos humanos; acceso a servicios esenciales públicos o privados –incluye riesgos de crédito de personas naturales–; seguridad pública; control de fronteras y migraciones; administración de la justicia.

En el caso de ser clasificados como sistemas de riesgo alto deberán incorporar una serie de requisitos básicos –sistema de identificación y gestión de riesgos; alta calidad de bases de datos; mantenimiento de registro automático de eventos; supervisión humana– y cumplir con unas obligaciones reforzadas –sistema de gestión de calidad; elaboración y transparencia de la documentación técnica del sistema; sistema de evaluación de conformidad; protocolos ante una no conformidad y proceso de notificación; marca CE de conformidad; entre otras–. Los importadores de tecnologías y sistemas de alto riesgo deberán garantizar que el sistema ha superado las pruebas de conformidad, que la documentación técnica ha sido redactada y publicada, que el sistema contiene la marca de conformidad requerida más la documentación y las instrucciones de utilización adecuadas y, en definitiva, que tanto el proveedor[17] como el importador del sistema han cumplido con las obligaciones establecidas en el Reglamento.

La tercera categoría de sistemas de inteligencia artificial, los de riesgo limitado, están regulados en el Título IV de la Propuesta –art. 52–, y comprende básicamente ciertas aplicaciones que interactúan con personas y que se emplean para detectar emociones o realizar asociaciones mediante categorías basadas en datos biométricos –como es característico de los robots conversacionales o de los sensores para la predicción y prevención de procesos críticos–, o que son susceptibles de generar o manipular contenido, como en el caso de los llamados *deep fakes,* que vinculan la imagen y voz de una persona con un mensaje que esta nunca llego a transmitir. Con respecto a estos sistemas de riesgo limitado, la normativa propuesta únicamente impone ciertos requisitos de transparencia, exigiendo que los sistemas de inteligencia artificial que interactúen con humanos deban informar del hecho que estos están hablando

17 A efectos de la propuesta de Reglamento, cualquier distribuidor, importador, usuario o tercero será considerado un proveedor cuando este comercialice u opere un sistema de alto riesgo bajo su propio nombre o marca; modifique el propósito de utilización del sistema; modifique de manera substancial el sistema.

y relacionándose con un algoritmo, explicando el nivel de exposición de las personas a la identificación de emociones o categorización biométrica y la eventual manipulación de imágenes, audio o video.

Los sistemas de inteligencia artificial que no se hallan comprendidos en ninguna de las tres categorías anteriores, se consideran de riesgo mínimo o nulo, de modo que no son objeto de regulación específica. Se indica empero que la Comisión Europea y los Estados miembros deberán estimular la aplicación voluntaria del Reglamento por parte de los proveedores de sistemas de bajo riesgo, mediante la adopción de códigos de conducta, que pueden ser establecidos por proveedores individuales o por organismos representativos de esa categoría.

La propuesta de Reglamento, en definitiva, establece un modelo de regulación basada en la responsabilidad proactiva de los distintos actores que participan en el desarrollo, implementación y comercialización de las herramientas basadas en inteligencia artificial, y lo hace con un enfoque que parte de la delimitación de niveles de riesgo en función de la tecnología empleada y de sus usos posibles[18].

Se fijan así una serie de líneas rojas con la categoría de riesgos inadmisibles, aquello que bajo ningún concepto estamos dispuestos a aceptar por su potencial invasivo o lesivo desde la óptica de los derechos fundamentales, y también un amplio abanico de soluciones de riesgo alto, que deberán cumplir con requisitos y obligaciones específicas muy diversas y exigentes.

IV. PROPUESTA SISTEMATIZADORA DE CLASIFICACIÓN DE LAS GARANTÍAS DE LOS SISTEMAS DE INTELIGENCIA ARTIFICIAL

Como hemos defendido más arriba los requisitos y obligaciones reforzadas de los sistemas de inteligencia artificial de riesgo alto actúan a modo de garantías «sistémicas» –siguiendo la terminología propuesta por la DSA–, con un enfoque integral y amplio –privacidad, seguridad, transparencia, responsabilidad,

18 La propuesta de Reglamento también ha sido objeto de estudio y análisis crítico en Simón (2022a: 143-153).

etc.– que encuentra su fundamento en la dimensión objetiva de los derechos fundamentales, la dignidad humana y el libre desarrollo de la personalidad, con el fin de reducir la incidencia o el nivel de impacto residual sobre los derechos fundamentales[19]. Esas garantías aplicadas en la práctica permiten a través de los controles mitigar los niveles de riesgo residual que se atribuyen a un concreto algoritmo tras una autoevaluación de impacto[20].

No es posible ni plausible tratar de ofrecer un listado de garantías que se apliquen a cualquier tipo de inteligencia artificial, sin tener en cuenta los usos concretos. Depende de cada tipo de tecnología y de su definición de uso, teniendo en cuenta el contexto, la naturaleza y el alcance de la herramienta o sistema. Así, por ejemplo, en el contexto de los sistemas inteligentes, la relevancia de las explicaciones suele depender del contexto y el tipo de aplicación de inteligencia artificial que se utilice, no siendo estas siempre estrictamente necesarias[21]. Otro buen ejemplo es el caso de la falaz medida de acceso al código íntegro, que podría clasificarse como una garantía dentro de la categoría de medidas que en ocasiones son útiles para garantizar la transparencia, pero que no es aplicable por inidónea en caso de algoritmos de aprendizaje profundo –*machine learning*– o en cajas negras con aplicación de redes neuronales.

Sin embargo, lo anterior no significa que debamos renunciar a tratar de sistematizar las garantías frente a la realidad algorítmica, y a continuación formularemos una propuesta de clasificación basada en las siguientes categorías: publicidad y transparencia; explicabilidad; seguridad y garantías sistémicas; garantías subjetivas y en la participación humana –diseño, entreno, implementación y monitorización–; garantías institucionales.

Existe una clara conexión entre todas las categorías propuestas, más aún si tenemos en cuenta que estamos hablando de garantías sistémicas; sin embargo, a nuestro modo de ver, cada una de las categorías indicadas tiene

19 Sobre este particular véanse Simón (2021: 161-202); Presno (2022).

20 De nuevo, interesa recordar que «el modelo de la gestión del riesgo, la responsabilidad proactiva y el diseño para el cumplimiento normativo tiene singular relevancia en el ámbito de la IA y el big data». Cotino (2022a: 93).

21 Véase, sobre la relevancia relativa de las explicaciones en el ámbito de la salud y atención sanitaria, el trabajo de Markus, *et al.* (2021).

autonomía conceptual suficiente como para ser objeto de estudio por separado, ubicándose en una escala jerárquica en posición de igualdad. Proponemos así la estructura propia de una taxonomía, basada en la jerarquía y formada por categorías y subcategorías, existiendo una relación de hermanos entre categorías y de padre-hijo entre las categorías y subcategorías.

IV.1. Transparencia algorítmica

Cuando hablamos de la transparencia de los algoritmos nos referimos a la publicidad activa y al derecho de acceso a la información relativa al funcionamiento técnico y, más concretamente, la que hace referencia a los propósitos y fines, a la estructura y diseño, a las acciones subyacentes, a las bases de datos empleadas y a los mecanismos de participación humana que estos incorporan. Esa información publicitada a terceros permite mantener la confianza de los usuarios del sistema y los capacita para impugnar sus resultados. No es una definición pacífica y otros autores definen la transparencia como aquella característica que convierte a un sistema inteligente en comprensible[22], lo que es difícil de aceptar entre otras cuestiones porque, como se verá a continuación, la relación entre transparencia y comprensibilidad no es necesariamente causal ni condicional. En una línea muy parecida a la definición de transparencia que acabamos de proponer, encontramos la concreción del extinto Grupo de Trabajo sobre Protección de Datos del Artículo 29, actual Comité Europeo de Protección de Datos (en adelante, CEPD), que en 2018 definió la transparencia como ese parámetro que permite «generar confianza en los procesos que afectan al ciudadano capacitándolo para entender y, en su caso, impugnar dichos procesos»[23]. Como se observa, el CEPD atribuye a la transparencia, como en la definición que proponemos, la capacidad de generar confianza y a su vez lo conecta con el derecho cívico de impugnar las decisiones automatizadas.

22 Aunque no es menos cierto que en esta propuesta la transparencia no deja de ser una arista más de un modelo de inteligencia artificial interpretable. Véase Lipton (2018).

23 Véase Grupo de Trabajo sobre protección de datos del artículo 29. «Directrices sobre la transparencia en virtud del Reglamento (UE) 2016/679». Adoptadas el 29 de noviembre de 2017, y actualizadas el 11 de abril de 2018. El CEPD ha hecho suyas estas directrices en virtud del *Endorsement* 1/2018, de 25 de mayo de 2018.

El principal problema en torno a la definición exacta del concepto de transparencia tiene que ver con la proximidad del mismo con otros términos[24], como la explicabilidad, categoría hermana, si bien es fácil encontrar literatura especializada que aceptando la relación más que evidente entre ambos conceptos, aboga por un uso distinto, teniendo en cuenta que esta última, como se verá más adelante, lo que realmente pretende es explicar o presentar sistemas de inteligencia artificial en términos comprensibles para los seres humanos[25].

Repárese ya en este momento que no es lo mismo entender o comprender como funciona un sistema de inteligencia artificial, gracias a las explicaciones ofrecidas en términos lógicos humanos, que el hecho que de forma activa se publicite y se garantice el acceso a una información determinada sobre el diseño, entreno, implementación y monitorización de este. En ocasiones, la transparencia es necesaria para poder explicar o comprender como funciona un sistema, pero su relación, en una y otra dirección, no es siempre de causalidad y tampoco actúa como condición. Ambas son empero categorías específicas que pueden intervenir en determinados procesos como garantías efectivas para alcanzar una inteligencia artificial confiable o ética, que se atomizan a su vez en un haz de subcategorías que definiremos y clasificaremos a continuación. Así las cosas, la lógica subyacente de la transparencia es incrementar la confianza y permitir el recurso o impugnación de sus decisiones con base a información –también los detalles técnicos– publicitados *ex ante*[26], de tal modo que el llamado «derecho a la explicación de las inferencias razonables» (Wachter, 2019: 572) se ubica dentro de la transparencia y no de la explicabilidad, pues sustituye la explicación en el caso concreto por una prueba previa de razonabilidad de la inferencia (Roig, 2020: 75). En cambio, la explicación de una decisión algorítmica se refiere a razones o justificaciones para un resultado concreto o particular, y no así del proceso

24 Sobre las dificultades semánticas y la necesidad de consensuar una definición de estas categorías, aunque en relación con el término explicabilidad, véase el trabajo de Ortiz (2022: 334).

25 Seguimos la definición propuesta por Doshi-Velez (2017).

26 Sobre los factores para tener en cuenta *ex ante* –explicación local– véase Roig (2020: 203).

de decisión en general, resultando las primeras innecesarias para impugnar o recurrir frente al sistema[27].

Las subcategorías de la transparencia[28] siguiendo el planeamiento indicado incluyen (1) la simulabilidad, (2) la descomponibilidad, (3) la legibilidad, (4) la auditabilidad y sus derivadas –auditado, auditable y verificable– y (5) la publicidad activa de los resultados y proyecciones, de los tests y comprobaciones.

Cuando hablamos de simulabilidad (1) hacemos referencia a la capacidad de un sistema de inteligencia artificial de ser simulado o pensado estrictamente por un ser humano, o lo que es lo mismo, que el modelo sea lo suficientemente autónomo para que un ser humano pueda pensar y razonar sobre este como un todo (Tulio, 2016). Los modelos lineales dispersos son más transparentes e interpretables que los densos. Los sistemas basados en reglas simples pero extensos quedarían fuera de esta clasificación y por ende son más opacos por naturaleza. En cambio, los sistemas basados en redes neuronales, por defecto más potentes, sí pueden ser en algunos casos incluidos dentro de los modelos simulables (Barredo, 2020: 87), en la medida que la red neuronal este compuesta por un solo perceptrón –una unidad de red neuronal–.

Un modelo descomponible (2) es aquel en el que cada una de sus partes –datos de entrada, parámetros y operaciones cálculo– es susceptible de ser comprendida por un ser humano sin necesidad de herramientas adicionales. La descomposición facilita la transparencia en la medida que la información relativa a cada una de sus partes es accesible y comprensible, si bien como sucedía anteriormente con (1) la simulabilidad, no todos los sistemas algorítmicos cumplen tampoco esta propiedad, como los de aprendizaje autónomo; de hecho, en ocasiones es posible separar sus partes, pero seguirá sin ser descomponible mientras una de sus partes no sea comprensible sin herramientas adicionales. Los algoritmos basados en sistemas de regresión logística y lineal,

27 La doctrina se ha referido explícitamente a las limitaciones de la transparencia para alcanzar un ideal de responsabilidad, si no es acompañada de la explicabilidad (Ananny, 2018).

28 No seguimos la misma taxonomía ni todos los atributos que otros autores han señalado previamente, aunque algunas de las subcategorías son compartidas por la doctrina técnica especializada. Véase al respecto Barredo (2020).

o los basados en los tradicionales árboles de decisión, sí cumplen con tal propiedad (Barredo, 2020: 90).

La legibilidad (3) es una característica que tiene que ver con la capacidad humana para leer predictores y variables del algoritmo, lo que sólo es posible si el modelo de inteligencia artificial no altera sus fórmulas o los datos que lo nutren, preservando su legibilidad. Se trata de un parámetro interrelacionado, como hermano, con las subcategorías (1) y (2), siendo una propiedad inexistente en los algoritmos que aplican redes neuronales y los basados en aprendizaje profundo.

La auditabilidad (4) es otra característica de la transparencia, aunque también forma parte de las tradicionales medidas de seguridad de los sistemas de información. Se trata de una propiedad que debe adscribirse parcialmente a la transparencia, únicamente por lo que se refiere a la publicidad activa de los informes de auditoría sobre el código algorítmico, y que además tiene diferentes niveles de intensidad: no es lo mismo que el código sea auditado, que en cambio sea auditable[29]. Por defecto, en las autoevaluaciones de riesgo y en caso de sistemas de riesgo alto, debería surgir la iniciativa por parte del desarrollador o responsable de la herramienta de auditar el código y presentar los resultados de las auditorías, que deberán ser recurrentes en el tiempo, aunque la periodicidad –anual, bianual, etc.– dependerá de factores que derivan del contexto, naturaleza y alcance de la solución tecnológica concreta. Que sea auditable es más difícil, incluso cuando los costes corran a cargo de aquél que lo solicita, puesto que estos procesos suponen también un coste de oportunidad –destinar tiempo y empleados para justificar procesos– nada desdeñable para la empresa u organismo público. Por este motivo debería descartarse en la gran mayoría de casos concretos, al ser desproporcionada en cuanto a costes para los responsables e incluso para el propio interesado; cuando la transparencia en realidad exige ese deber proactivo publicitando los resultados de las auditorias previas. De auditorías, además, hay de muchos tipos en función de la metodología empleada: auditar el código analizando el programa y las bases de datos usadas para el entrenamiento; auditoría no invasiva por el usuario; método *sock puppet*; auditoría colaborativa; auditoría

29 Las auditorías implican llevar al límite, de nuevo, al algoritmo, y realizar un estudio específico a la búsqueda de los errores de este. Véase De Laat (2018: 538).

scraping analizando únicamente el entrenamiento con pruebas de estrés del sistema[30]. Uno u otro modelo serán más o menos aplicables en relación con los resultados del análisis previo del contexto, alcance y naturaleza de la tecnología aplicada al supuesto objeto de estudio.

Sucede algo parecido a (4) la auditabilidad, cuando analizamos la (5) publicidad activa de los tests, pruebas y entrenamiento del algoritmo. No debe, el responsable de la herramienta de inteligencia artificial, publicitar o informar de todo, descartando lo accesorio e incorporando únicamente lo relevante, desde la óptica cívica del usuario que, en el futuro, pueda querer reclamar o impugnar las proyecciones o decisiones del algoritmo. De nuevo, se trata de una medida de seguridad tradicional de los sistemas de información, pero la publicidad de los resultados integra necesariamente la categoría de la propiedad de la transparencia. Se puede emplear como término para referirse a esta propiedad, indistintamente, la verificabilidad, testabilidad o comprobabilidad del sistema.

La transparencia, con todo, es una garantía relativa, que debe leerse en cada caso concreto y en relación con las otras propiedades descritas en la taxonomía de las garantías de los sistemas de inteligencia artificial. Para algunos autores, la transparencia no es suficiente frente a determinadas realidades algorítmicas, puesto que esa información que se publicita para generar confianza y la posibilidad de recurrir se ven condicionadas por las dificultades de comprender los resultados concretos, lo que nos exige entrar en el campo de la explicabilidad[31].

IV.2. Explicabilidad

Hemos diferenciado anteriormente la explicabilidad de la transparencia, pero ante la falta de consenso, y ante la limitación de operar sólo con una definición en términos comparativos de la citada categoría, procede matizar que nos

30 Para más detalles sobre las distintas metodologías de auditorías algorítmicas véase Bernhard (2018: 614).

31 Al respecto, Roig señala que «en las decisiones automatizadas no parece necesario detallar los aspectos técnicos del algoritmo, pero sí en cambio se pueden explicar los factores que se han tenido en cuenta, así como sus consecuencias para el interesado» (2020: 48).

referimos a aquella propiedad que hace inteligible los resultados de los sistemas de inteligencia artificial en un caso concreto, así como la comprensibilidad de los datos, procesos y comportamientos asociados a la decisión específica que se proyecta sobre los individuos[32]. La explicabilidad actúa como garantía en la medida que permite hacer comprensible, entendible o inteligible la aplicación individualizada de un algoritmo a un supuesto de hecho específico, justificando la racionalidad o criterios que hay detrás de una decisión (Ortiz, 2022: 334).

Las subcategorías que integran la explicabilidad son (1) la inteligibilidad, (2) la comprensibilidad y (3) la interpretabilidad. La primera de ellas (1) es la característica de un modelo que permite al ser humano comprender la función –cómo funciona el sistema– del algoritmo, sin necesidad de explicar su estructura interna o el modelo que le permite procesar los datos internamente (Montavon, 2018). La comprensibilidad (2) exige que el sistema de inteligencia artificial represente y explique el conocimiento aprendido de una forma comprensible para los humanos. Dada su difícil cuantificación, la comprensibilidad normalmente está ligada a la evaluación de la complejidad del sistema (Fernández, 2019). La interpretabilidad (3) se relaciona con la habilidad de explicar u ofrecer el significado de la decisión adoptada en términos comprensibles para un humano. Tal propiedad permite, en realidad, garantizar la imparcialidad en la toma de decisiones, por ejemplo, al detectar y consecuentemente corregir sesgos en el entrenamiento con el conjunto de datos (Barredo, 2020: 83). La detección se produce en el proceso en el que propio algoritmo trata de explicar u ofrecer el significado de la decisión adoptada. O lo que es lo mismo, la posibilidad de explicar los mecanismos causales de la inteligencia artificial también posibilita la resolución de problemas de este a nivel técnico (Athey, 2015). Dentro de este orden de ideas, la explicabilidad en sentido estricto se asocia a la creación de una interfaz entre humanos y quien toma la decisión, en la que el segundo trata de explicar el significado y las razones que subyacen a la decisión concreta adoptada.

En consecuencia, se puede inferir que es mucho más difícil y complejo que un sistema algorítmico contenga la propiedad de la explicabilidad,

32 Coincide en lo sustantivo, aunque no es literal, con la definición de explicabilidad propuesta por UNESCO. (2021). «First draft of the recommendation on the ethics of artificial intelligence». Disponible en Internet: https://unesdoc.unesco.org/ark:/48223/pf0000373434 (última consulta el 9 de julio de 2022).

especialmente, si la comparamos con la transparencia. Por ello se ha señalado que ante las limitaciones y dificultades prácticas para garantizar la explicabilidad, esta debería ser equilibrada con la transparencia y con un sistema en el que, siguiendo a Medina, las evaluaciones de impacto juegan un papel relevante para defender un sistema de «legibilidad por diseño»[33]. En función de la tecnología aplicada y del uso concreto, la explicabilidad no es una alternativa real, por inviable (Edwards, 2017: 65), y en esos casos habrá que apoyarse en otras garantías en función de los resultados de las autoevaluaciones de impacto[34].

Que en muchas ocasiones la explicabilidad sea imposible de alcanzar, en la práctica, tampoco debe significar una renuncia anticipada a la pretensión de que algunas herramientas tecnológicas intenten o traten de reunir esa cualidad. Además, cabe recordar que la explicabilidad es una propiedad que contribuye a crear una falsa sensación de seguridad (Russell, 1995) que actúa a modo de lastre contra la innovación tecnológica. La falacia se produce en la medida que finalmente se acaban escogiendo tecnologías menos seguras, ricas o efectivas por el mero hecho de que permiten ser explicadas. Por ejemplo, cuando para un uso determinado preferimos construir un sistema experto que emule los tradicionales árboles de decisión humana, rechazando aplicar redes neurales, con lo que condenamos el modelo a las falibles fórmulas decisorias humanas.

IV.3. Seguridad

Las tradicionales medidas de seguridad contempladas en los estándares y buenas prácticas relativas a los sistemas de gestión de la seguridad de la información integran, en bloque, una de las principales propiedades de la taxonomía de las garantías de los sistemas de inteligencia artificial. Que el desarrollo tecnológico debe ser seguro significa que la base de datos que nutre el algoritmo y la información clave relacionada con los componentes

33 Como se ha detallado anteriormente, la legibilidad constituye una subcategoría de la transparencia algorítmica. Sobre el papel de las evaluaciones de impacto «para superar el insatisfactorio resultado alcanzado con el derecho a la explicación», véase Medina (2022: 170).

34 Lo importante «no es tanto el reconocimiento de un sólido derecho a la explicación como avanzar hacia un sistema de algoritmos interpretables». Medina (2022: 170).

y procesos de este deben cumplir con un estándar elevado de protección, lo que significa encriptación en viaje y descanso de la información, el cifrado de extremo a extremo, la seguridad en el diseño y por defecto durante todo el ciclo de vida de la herramienta tecnológica, garantizar la confidencialidad –que también se ha conectado a los problemas sobre la propiedad del algoritmo y los secretos de empresa[35]–, integridad y disponibilidad de la información relacionada y los datos empleados, y quizás lo más importante, por tener una aplicación transversal, y conectada también con la transparencia y las garantías subjetivas, asegurar la trazabilidad de cualquier proceso o acción. Esto significa aplicar todas las medidas, garantías y salvaguardas pensadas para los sistemas de tecnologías de la información a los sistemas de inteligencia artificial[36].

La trazabilidad (1) no es propiamente una subcategoría o propiedad de la seguridad, puesto que no podemos hablar de seguridad sin referimos necesariamente a la trazabilidad de los procesos y acciones sobre el algoritmo, desde la idea y diseño hasta la última actualización o despliegue –*deploy*, siguiendo la terminología empleada en los entornos de desarrollo tecnológico– de este. Sin trazabilidad, entonces, no podemos hablar de un sistema seguro, resultando por ende esta una propiedad clave y transversal, habitual como decíamos anteriormente de los sistemas de seguridad de la información[37].

Para poder hablar de trazabilidad, cualquier acción que lleve a cabo un usuario del sistema –con indiferencia de si es el propietario o administrador principal– quedará registrada y dejará un rastro que, en el futuro y en el caso de ser necesario, podrá ser examinada[38]. Se trata de una función en entornos en línea relacionada con los *backlogs* –registros con las acciones que cada usuario realiza dentro de un sistema, desde el mero acceso hasta su conexión– o los llamados *audit trail* –pistas para auditoría, como traducción literal–.

35 Véase el trabajo de Adadi (2018: 52138-52160).

36 En la misma dirección véase Barredo (2020: 108).

37 Véanse por ejemplo las normas (buenas prácticas) UNE-EN ISO/IEC 27001:2017 y ISO/IEC 27005:2018.

38 Para más detalles véase el interesante trabajo de Kroll (2021: 758-771).

Como es lógico, la existencia de estos registros puede resultar fundamental para poder llevar a cabo con éxito auditorias del código, aunque debe soslayarse con claridad de (2) la auditabilidad, como propiedad, que también es una medida de seguridad, aunque sus resultados e informes, como decíamos anteriormente, formen parte de la transparencia. Las auditorias son una pieza clave de los sistemas de seguridad de la información, al comprobar, revisar y redefinir, dentro del marco del principio de mejora continua, la efectividad de los controles y la salud del sistema en la práctica, con propuestas de acciones de mejora a futuro[39]. La trazabilidad ayuda a estos procesos puesto que permite comprobar si se han producido errores en el pasado que puedan ser evitados en el futuro, y su carácter transversal hace que se establezcan conexiones con la transparencia y la explicabilidad –en lo relativo a la posibilidad de explicar el razonamiento que ha llevado a tomar una decisión en un caso concreto–.

Sea como fuere, parece evidente que este tipo de medidas de seguridad específicas, propias de sistemas de gestión de la seguridad de la información, deben estar presentes más aún cuando estamos hablando de sistemas de riesgo alto que emplean tecnologías de vanguardia con un potencial lesivo significativo.

IV.4. La participación o intervención humana

Definir con precisión quién es –el sujeto– responsable del diseño, desarrollo, implementación y monitorización del sistema de inteligencia artificial proyecta también sus efectos sobre los derechos en juego y, por ende, una serie de garantías provenientes de normas de derecho fuerte deberían asegurar una mínima homogeneidad que en la práctica suponga que los ciudadanos son capaces de impugnar la decisión algorítmica[40].

El legislador europeo avanza hacia un modelo de corregulación más que de autorregulación –a diferencia de los «mecanismos de autorregulación

39 Véanse las normas UNE-ISO citadas anteriormente.

40 En relación con la aplicación de los sistemas de inteligencia artificial en la Administración de Justicia, y sobre la importancia de definir con precisión el quién –órgano de control–, véase Simón (2021: 203-207 y 209-213).

transparentes» a los que se refiere la Carta de derechos digitales[41]–. Esto implica que las empresas desarrolladoras deben actuar *ex ante* con autoevaluaciones de riesgo, pero parece poco prudente que, ante los sistemas de alto riesgo –siguiendo la terminología y la categoría establecida en la propuesta de reglamento europeo–, cuando los algoritmos se empleen por parte del sector público para la toma de decisiones automatizadas, se confíe ciegamente en esa proactividad del sector privado sin exigir algo más que un sello de conformidad del regulador europeo. En este sentido, podrían establecerse condiciones de *hard law* que exijan, en función del caso concreto, el contexto y la naturaleza, o bien la compra o diseño propio de la solución tecnológica por parte del organismo público en cuestión –ideal, aunque inalcanzable en la mayoría de las ocasiones por incapacidad presupuestaria– o bien el control relativo de ciertos procesos, como los relacionados con la monitorización, auditoria, trazabilidad y mejora continua del sistema.

No procede aquí, por limitaciones espaciales evidentes y por no tratarse del objeto del presente trabajo, hacer un estudio de *lege lata et ferenda* sobre qué controles subjetivos –relativos a la participación de una persona física o jurídica, empresa u organismo público– sería más adecuado que el legislador europeo o nacional establecieran *ex lege*, y mucho menos la determinación de a partir de qué nivel de riesgo residual es necesario exigir esas garantías reforzadas en función del sujeto. Repárese que ciertos problemas no sólo aparecen cuando una Administración pública emplea un algoritmo cuya propiedad es de una empresa privada, sino que también cuando una empresa privada u organismo público confía en una tecnología que ha sido desarrollada por un tercero ubicado en un país extranjero que no ofrece garantías equivalentes a las que exigirá, en un futuro cercano, la normativa europea. Se trata de dos parámetros siempre presentes en las evaluaciones de riesgo, las variables de externalización –dependencia de terceros– y el riesgo en la ejecución por parte de estos –*third party risks*, siguiendo la terminología empleada en entornos de cumplimiento normativo–, y que deben tenerse en cuenta para el cumplimiento de la normativa.

41 Nos referimos a la Carta de Derechos Digitales, que no tiene valor normativo, disponible en Internet: https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf (última consulta el 9 de julio de 2022).

Si podemos empero formular una premisa que bien podría incorporarse en las normas que pretenden regular los sistemas de inteligencia artificial con vocación armonizadora, del tenor siguiente: «en el caso que el análisis, fruto de una autoevaluación de riesgos sistémicos en el ámbito de sistemas de inteligencia artificial de la categoría de riesgo alto, arroje un riesgo residual concreto por encima del umbral aceptable, deberán implementarse una serie de controles específicos reforzados en relación con el sujeto –empresa privada u organismo público– encargado del diseño, entrenamiento, implementación y monitorización del algoritmo».

Esos controles podrían ser, como decíamos anteriormente, el desarrollo exclusivo *inhouse,* la compra de la solución tecnológica o la atribución por parte del propietario de poderes concretos de control –en los procesos de diseño, auditoría, revisiones del sistema con acceso a los *backlogs* que permiten la trazabilidad, entrenamiento y monitorización– a la Administración pública, organismo público o empresa privada que empleará el algoritmo en el caso concreto.

No es una cuestión estrictamente novedosa si la comparamos con aquello que ya ha sucedido en relación con la evaluación de impacto en protección de datos, con la intervención de autoridades de control y no del legislador, que han limitado el margen de discrecionalidad en la valoración de la necesidad de llevar a cabo esta evaluación específica –con *whitelists* y *blacklists*– y en la definición o propuesta de controles específicos[42]. Aquí se trataría de ir un poco más allá, estableciendo por Ley la exigencia de controles reforzados ante determinados niveles de riesgo residual, a las que venimos refiriéndonos como salvaguardas subjetivas. Con todo, consideramos importante plantear este extremo dentro de la taxonomía de las garantías de los sistemas de inteligencia artificial, haciendo hincapié en la necesidad de intervención futura por parte del legislador.

IV.5. Garantías institucionales

Los derechos fundamentales gozan de garantías institucionales concretas como la figura del Defensor del Pueblo, y además, ciertos derechos –como la

42 Véanse por ejemplo las listas de los tipos de tratamiento –que requieren o no– realizar una evaluación de impacto en protección de datos. Disponibles en Internet: https://www.aepd.es/es/documento/listasdpia-35.5l.pdf y https://www.aepd.es/es/documento/listas-dpia-es-35-4.pdf (última consulta el 9 de julio de 2022).

protección de datos personales– disponen incluso de una autoridad independiente de control, cuya finalidad primordial no es otra que velar por el cumplimiento de la legislación sobre protección de datos y controlar su aplicación, desempeñando funciones de inspección y sanción, así como una tarea de estudio y elaboración de guías, instrucciones o recomendaciones relacionadas con el citado derecho.

¿Es necesaria una autoridad independiente supervisora garante de la gobernanza algorítmica, de forma parecida a lo que sucede con la protección de datos? La realidad es que parece imponerse la idea, tanto a nivel doctrinal (Roig, 2021: 237-238). como a nivel normativo –en perspectiva comparada, como se verá a continuación–, que ante unos niveles de riesgo residual elevados es necesario que una autoridad especializada, con un enfoque específico en los principios que pretenden alcanzar una inteligencia artificial ética y confiable, ejerza funciones, más allá de su participación en los sellos de conformidad o en los hipotéticos futuros esquemas de certificación, con cometidos de inspección, sanción y elaboración de recomendaciones, avisos y guías sobre la materia.

El *Institute of Electrical and Electronics Engineers* (en adelante, IEEE), una asociación profesional internacional dedicada a la estandarización, en su iniciativa global para el desarrollo ético de los sistemas autónomos e inteligentes, plantea como problema la falta de una organización independiente de revisión «para supervisar si tales productos realmente cumplen criterios éticos, tanto cuando son desplegados, como considerando su evolución tras el despliegue e interacción con otros productos» (Shahriari, 2017), y advierte de la «brecha entre cómo se comercializan los sistemas de inteligencia artificial y su desempeño real o aplicación. Necesitamos asegurarnos de que la tecnología va acompañada de las mejores recomendaciones de uso y advertencias asociadas. Además, necesitamos desarrollar un esquema de certificación para los sistemas que asegure que las tecnologías han sido evaluadas de forma independiente como seguras y éticamente sólidas» (Shahriari, 2017).

La creación de un modelo basado en la corregulación y las autoevaluaciones de riesgo necesita, en paralelo, de una autoridad que revise y mantenga el registro –con evidencias– de la conformidad de los productos, y que establezca un esquema en base a los estándares, convenciones y normativas vigentes que permita la certificación voluntaria de los anteriores. Además,

esta autoridad podría actuar como incentivo al cumplimiento de empresas privadas –aunque en sentido negativo, bajo amenaza de inspección y sanción– para que el desarrollo tecnológico sea plenamente respetuoso con los principios y garantías de un sistema de inteligencia artificial ético y confiable, jugando también un papel significativo desde el punto de vista del ejercicio de funciones de coordinación de los órganos o autoridades nacionales de los distintos Estados miembros y de asesoramiento e informativas, aportando cierta seguridad jurídica con la elaboración de guías sectoriales, recomendaciones o avisos.

En perspectiva comparada, proliferan organismos con competencias sobre la materia, aunque su naturaleza, estructura y funciones no son comparables con la propuesta que incorpora el título VI de la propuesta de reglamento europeo sobre inteligencia artificial, que prevé la constitución del Comité Europeo de Inteligencia Artificial. En Reino Unido, por ejemplo, se ha constituido el *Centre for Data Ethics and Innovation* que se encarga de «conectar la confianza en el uso de datos y la inteligencia artificial»[43]; en Canadá encontramos el *Advisory Council on Artificial Intelligence*[44] que, integrado por grupo de expertos en la materia, se encarga de asesorar al Gobierno canadiense para establecer una estrategia global de liderazgo en el sector, identificando oportunidades y tratando de minimizar el impacto y los riesgos de los sistemas de inteligencia artificial, con políticas, recomendaciones e informes anuales.

Sin embargo, como decíamos, no es lo mismo un consejo de expertos, con capacidad de seguimiento, de emitir informes y de asesoramiento, que un organismo con atribución de funciones relacionadas con la inspección, la imposición de sanciones, la emisión de sellos de conformidad y la llevanza de su registro, la coordinación de autoridades nacionales de control, etc.

La citada propuesta de reglamento europeo, en el art. 59, se refiere a la designación de las autoridades nacionales competentes que, en realidad, se configura como un mandato u obligación preceptiva para los Estados

43 Puede consultarse en Internet: https://www.gov.uk/government/organisations/centre-for-data-ethics-and-innovation (última consulta el 9 de julio de 2022).

44 Puede consultarse en Internet: https://ised-isde.canada.ca/site/advisory-council-artificial-intelligence/en (última consulta el 9 de julio de 2022).

miembros, con el fin de garantizar la aplicación y ejecución del reglamento, exigiendo además que preserven la objetividad e imparcialidad de sus actividades y funciones. Se establece en el cuarto epígrafe del citado artículo una obligación de medios –recursos financieros y humanos adecuados– y la necesaria perspectiva multidisciplinar del grupo de personas que integren los equipos de la autoridad nacional de control, y en el quinto, la obligación de estas de presentar un informe anual a la Comisión con una evaluación de la idoneidad de los recursos financieros y humanos de las autoridades nacionales competentes.

En España, ya disponemos de un consejo asesor[45] dotado de plena autonomía funcional desde julio de 2020, con funciones limitadas a asesorar e informar al Ministerio de Asuntos Económicos y Transformación Digital, valorar observaciones y comentarios, así como formular propuestas sobre la Estrategia Nacional de Inteligencia Artificial y asesorar en materia de evaluación de impacto en la industria, la Administración y la sociedad.

Por otro lado, la intención del Gobierno es crear, antes que ningún otro Estado miembro, una autoridad nacional de control siguiendo las previsiones de la propuesta europea de reglamento de inteligencia artificial. La citada autoridad ya ha sido bautizada como la «Agencia Española de Supervisión de la Inteligencia Artificial», y se incluye en la Ley de Presupuestos Generales del Estado[46] para el año 2022, en la disposición adicional centésima trigésima, que incorpora información sobre la naturaleza y funciones del futuro órgano, más concretamente, cuando nos dice que será una Agencia Estatal dotada de personalidad jurídica pública, patrimonio propio y autonomía en su gestión, con potestad administrativa, y que «actuará con plena independencia orgánica y funcional de las Administraciones Públicas, de forma objetiva, transparente e imparcial, llevando a cabo medidas destinadas a la minimización de riesgos significativos sobre la seguridad y salud de las personas, así como sobre

45 Creado y regulado por la Orden ETD/670/2020, de 8 de julio, por la que se crea y regula el Consejo Asesor de Inteligencia Artificial. Publicado en el BOE núm. 199, de 22 de julio de 2020, pp. 55144-55147.

46 Véase la disposición adicional centésima trigésima de la Ley 22/2021, de 28 de diciembre, de Presupuestos Generales del Estado para el año 2022. Publicado en el BOE núm. 312, de 29 de diciembre de 2021, pp. 165114-165875.

sus derechos fundamentales, que puedan derivarse del uso de sistemas de inteligencia artificial»[47].

V. CONCLUSIONES

La tecnología es neutra[48], aunque en su desarrollo los seres humanos podemos proyectar sesgos y heurísticos, de tal modo que su empleo e implementación distribuya beneficios y perjuicios de manera desigual sobre los ciudadanos, sobre los que no deberían proyectarse decisiones automatizadas sin capacidad de reacción o recurso.

A lo largo de este trabajo hemos analizado la insuficiencia de un marco normativo atomizado para dar respuestas y regular el escenario tan ambicioso y complejo que dibujan los sistemas de inteligencia artificial. Europa quiere, de nuevo y como ya sucedió con el RGPD, ser pionera delimitando qué tecnología –con ciertas líneas rojas propias de *hard law*–, para qué –exigiendo un debate centrado en los usos éticos y plausibles–, quién –participación humana– y cómo –principios y garantías, incluida la creación de un órgano de control–.

En relación con las hipótesis planteadas en la introducción de este trabajo, hemos alcanzado las siguientes conclusiones:

Primera.- No es posible dar respuesta –principios, obligaciones, derechos, garantías– al desarrollo y adopción de sistemas de inteligencia jurídica artificial con base en la normativa europea de protección de datos, cuyo alcance es limitado e insuficiente. El desafío que plantea la realidad algorítmica requiere una lectura sistemática de los principios y derechos constitucionales, incluyendo un enfoque de riesgo propio de los sistemas de *compliance* –que ha mostrado su eficacia en otros ámbitos, como la prevención del delito–, adscribiendo la respuesta del Derecho frente a la tecnología a la dimensión objetiva de los derechos fundamentales y al papel reservado por la Constitución a la dignidad humana y al libre desarrollo de la personalidad.

47 *Ibídem.*

48 Una cosa bien distinta es que los desarrolladores proyecten sesgos en el código algorítmico.

Segunda.- Las garantías frente al uso de sistemas de inteligencia artificial deben clasificarse y sistematizarse mediante un esfuerzo de dogmática aplicada; es necesario conceptuar y sistematizar los factores y (sub) propiedades de garantías tales como la seguridad, la transparencia y la explicabilidad.

Tercera.- Debe superarse la confusión conceptual entre transparencia y explicabilidad, siendo garantías con autonomía conceptual y propiedades diferenciadas. La transparencia *per se* no es la panacea de todos los males ni garantiza necesariamente que un individuo no sea objeto de una decisión automatizada o que este se pueda defender frente a ella; al igual que sucede con las demás garantías o propiedades, cuya lectura aislada no aporta nada en el marco de las autoevaluaciones de impacto. El conjunto de las garantías estudiadas pretende un objetivo mayor: asegurar que ningún ciudadano sea objeto de una decisión automatizada sin poder recurrir, replicar o defenderse frente a tal decisión.

Cuarta.- De la premisa anterior se infiere la necesidad de establecer otras garantías como la seguridad –en especial, la trazabilidad como subpropiedad–, la participación humana –garantías subjetivas– o la supervisión *ex ante* y *ex post* –garantías institucionales–, todas ellas necesarias si lo que se pretende es alcanzar un ideal de IA confiable, responsable y ética.

VI. BIBLIOGRAFÍA

- ADADI, A. y BERRADA, M. (2018). «Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)». IEEE Access, vol. 6, 52138-52160.

- ANANNY, M. y CRAWOFRD, K. (2018). «Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability». *New media & society*, vol. 20, 3, 973-989.

- ATHEY, S. y IMBENS, G. (2015). «Machine Learning Methods for Estimating Heterogeneous Causal Effects». *arXiv*: 1504.01132.

- BARATA, J. (2021). «The Digital Services Act and its impact on the right to freedom of expression: special focus on risk mitigation obligations».

PLI, Plataforma por la Libertad de Información, disponible en https://libertadinformacion.cc/wp-content/uploads/2021/06/DSA-AND-ITS-IMPACT-ON-FREEDOM-OF-EXPRESSION-JOAN-BARATA-PDLI.pdf (última consulta el 9 de julio de 2022).

– BARREDO ARRIETA, A. *et al.* (2020). «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI». *Information Fusion*, 58, 82-115.

– BERNHARD WALT, R. V. (2018). «Increasing Transparency in Algorithmic Decision-Making with Explainable AI». *Datenschutz und Datensicherheit*, 10, 613-617.

– BOIX PALOP, A. (2020a). «Algorithms as Regulations: Considering Algorithms, when Used by the Public Administration for Decision-making, as Legal Norms in order to Guarantee the proper adoption of Administrative Decisions». *European Review of Digital Administration & Law*, Vol. 1, 1-2, 75-100.

- (2020b). «Los algoritmos son reglamentos: la necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones». *Revista de Derecho Público: Teoría y método*, Vol. 1.

– CERRILLO I MARTÍNEZ, A. (2020). «La transparencia de los algoritmos que utilizan las administraciones públicas». *Anuario de Transparencia Local*, 3, 41-78.

– COTINO HUESO, L. (2018). «La necesaria actualización de los derechos fundamentales como derechos digitales ante el desarrollo de internet y las nuevas tecnologías», en PENDÁS GARCÍA, B. (dir.), *España constitucional* (1978-2018): trayectorias y perspectivas, Vol. 3, Tomo 3, Madrid, Centro de Estudios Políticos y Constitucionales, 2347-2361.

- (2022a). «Nuevo paradigma en las garantías de los derechos fundamentales y una nueva protección de datos frente al impacto social y colectivo de la inteligencia artificial», en BAUZÁ REILLY, M. (Coord.)

y COTINO HUESO, L. (Dir.), *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Cizur Menor, Aranzadi, 69-105.

- (2022b). «Quién, cómo y qué regular (o no regular) frente a la desinformación». *Teoría y Realidad Constitucional*, 49, 199-238.

– COTINO HUESO, L. *et al.* (2021). «Un análisis crítico constructivo de la propuesta de Reglamento de la Unión Europea por le que se establecen normas armonizadas sobre la Inteligencia Artificial (Artificial Intelligence Act)». *Diario La Ley*.

– DE LAAT, P. B. (2018). «Algorithmic Decision-making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?». *Philos. Technol.*, 31, 525-541.

– DOSHI-VELEZ, F. y KIM, B. (2017). «Towards a rigorous science of interpretable machine learning». *arXiv preprint*: 1702.08608.

– EDWARDS, L. y VEALE, M. (2017). «Slave to the Algorithm? Why a 'Right to Explanation' is probably not the Remedy you are looking for». *Duke Law & Technology Review*, 16, 18-84.

– FERNÁNDEZ, A. (2019). « Evolutionary Fuzzy Systems for Explainable Artificial Intelligence: Why, When, What for, and Where to?». *IEEE Computational Intelligence Magazine*, 14, 1, 69-81.

– Gobierno de Canadá. (2022). «Algorithmic Impact Assessment». Disponible en Internet: https://open.canada.ca/aia-eia-js/?lang=en (última consulta el 9 de julio de 2022).

– Gobierno de Holanda. (2021). «Fundamental Rights and Algorithms Impact Assessment (FRAIA)». Disponible en Internet: https://www.government.nl/binaries/government/documenten/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms/fundamental-rights-and-algorithms-impact-assessment-fraia.pdf (última consulta el 9 de julio de 2022).

– JOBIN, A. *et al*. (2019). «The global landscape of AI ethics guidelines». *Nat Mach Intell*, 1, 389–399.

– KROLL, J. A. (2021). «Outlining traceability: A principle for operationalizing accountability in computing systems». *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transpare*ncy, 758-771.

– LIPTON, Z. C. (2018). «The Mythos of Model Interpretability». *Queue*, 16, 3.

– MADAIO, M. A. *et al.* (2020). «Co-designing checklists to understand organizational challenges and opportunities around fairness in AI». *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-14.

– MANTELERO, A. (2018). *El big data en el marco del Reglamento General de Protección de Datos*, Barcelona, UOC, 1-46.

- (2017). «From group privacy to collective privacy: towards a new dimension of privacy and data protection in the big data era», en TAYLOR, L. *et. al* (Eds.), *Group Privacy: New Challenges of Data Technologies*, La Haya, Springer, 173-198.

- (2022). *Beyond Data. Human Rights, Ethical and Social Impact Assessment in AI*, La Haya, Springer, 1-91 y 185-197.

– MARCHENA, M. (2022). «Inteligencia Artificial y jurisdicción penal». Discurso pronunciado en el acto de su toma de posesión como académico de número de la Real Academia de Doctores de España, 26 de octubre de 2022.

– MARKUS, A. F. et al. (2021). «The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies». *Journal of Biomedical Informatics*, 113.

– MEDINA GUERRERO, M. (2022). «El derecho a conocer los algoritmos utilizados en la toma de decisiones. Aproximación desde la

perspectiva del derecho fundamental a la protección de datos personales». *Teoría y Realidad Constitucional*, 49, 141-171.

– MIGUEL ASENSIO, P. A. (2021). «Propuesta de Reglamento sobre inteligencia artificial». *La Ley Unión Europea*, núm. 92.

– MIRÓ LLINARES, F. (2018). «Inteligencia artificial y justicia penal: más allá de los resultados lesivos causados por robots». *Revista de Derecho Penal y Criminología*, núm. 20, pp. 87-130.

- (2020). «Policía predictiva: ¿utopía o distopía? Sobre las actitudes hacia el uso de algoritmos de big data para la aplicación de la ley». IDP. *Revista de Internet, Derecho y Política*, núm. 30.

– MORLEY, J. *et al.* (2019). «From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices». *Science and Engineering Ethics*, 26, 4, 2141-2168, más concretamente pp. 2157 y ss.

– MONTAVON, G. *et al.* (2018). «Methods for interpreting and understanding deep neural networks». *Digital Signal Processing*, 73, 1-15.

– ORTIZ DE ZÁRATE ALCARAZO, L. (2022). «Explicabilidad (de la inteligencia artificial)». *Eunomía. Revista en Cultura de la Legalidad*, 22, 328-344, más concretamente, 334-335.

– PRESNO LINERA, M. A. (2022). «Una aproximación a la inteligencia artificial y su incidencia en los derechos fundamentales». *IDP: Observatorio de Derecho Público*, disponible en Internet: https://idpbarcelona.net/una-aproximacion-a-la-inteligencia-artificial-y-su-incidencia-en-los-derechos-fundamentales/ (última consulta el 4 de agosto de 2022).

– RALLO LOMBARTE, A. (2020). «Una nueva generación de derechos digitales». *Revista de Estudios Políticos*, 187, 101-135.

– RANCHORDAS, S. (2021). «Experimental Regulations for AI: Sandboxes for Morals and Mores». *University of Groningen Faculty of Law*

Research Paper, 7/2021, disponible en Internet: http://dx.doi.org/10.2139/ssrn.3839744 (última consulta el 9 de julio de 2022).

– ROIG BATALLA, A. (2020). *Las garantías frente a las decisiones automatizadas*, Barcelona, J. M. Bosch.

– RUSSELL, S. J. y NORVIG, P. (1995). *Artificial Intelligence: A Modern Approach*. New Jersey, Prentice Hall.

– SHAHRIARI, K. y SHAHRIARI, M. (2017). «Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems», *IEEE standard review*, p. 70, disponible en Internet: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf (última consulta el 9 de julio de 2022).

– SIMÓN CASTELLANO, P. (2021). *Justicia cautelar e inteligencia artificial: la alternativa a los atávicos heurísticos judiciales*, Barcelona, J. M. Bosch.

- (2022a). *La prisión algorítmica: Prevención, reinserción social y tutela de derechos fundamentales en el paradigma de los centros penitenciarios inteligentes*, València, Tirant lo Blanch.

- (2022b). «Nuevo Derecho, nuevas garantías. Una propuesta de reinterpretación de los principios jurídicos a la luz de la realidad algorítmica». *Revista de Derecho Político,* UNED, en prensa.

– SMUHA, N. A. *et al*. (2021). «How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act». Disponible en Internet: http://dx.doi.org/10.2139/ssrn.3899991 (última consulta el 9 de julio de 2022).

– TULIO RIBEIRO, M. *et al*. (2016). «Why Should I Trust You?: Explaining the Predictions of Any Classifier». KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

– UNESCO. (2021). «First draft of the recommendation on the ethics of artificial intelligence». Disponible en Internet: https://unesdoc.unesco.org/ark:/48223/pf0000373434 (última consulta el 9 de julio de 2022).

– WACHTER, S. y MITTELSTADT, B. (2019). «A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI». *Columbia Business Law Review*, 2, 494-620.