TAXONOMÍA DE LAS GARANTÍAS JURÍDICAS EN EL EMPLEO DE LOS SISTEMAS DE INTELIGENCIA ARTIFICIAL

PERE SIMÓN CASTELLANO

SUMARIO

Fecha recepción: 10.08.2022
Fecha aceptación: 7.01.2023

TAXONOMÍA DE LAS GARANTÍAS JURÍDICAS EN EL EMPLEO DE LOS SISTEMAS DE INTELIGENCIA ARTIFICIAL

PERE SIMÓN CASTELLANO[1]

Profesor Titular de Derecho Constitucional
Universidad Internacional de la Rioja – UNIR
Magistrado (suplente) de la Audiencia Provincial de Girona

I. A MODO DE INTRODUCCIÓN: NUEVO DERECHO, NUEVAS GARANTÍAS.

La doctrina nacional ha tenido la posibilidad de estudiar parcialmente la metamorfosis de los ordenamientos jurídicos como consecuencia de los desplazamientos producidos en sus elementos consecutivos, cuyo principal corolario no es otro que la existencia de un nuevo Derecho, por lo que se refiere incluso a su significado y garantía[2]. Dichos desplazamientos o cambios de orientación obedecen al protagonismo

---

[1] Profesor Titular de Derecho Constitucional. Universidad Internacional de la Rioja. Facultad de Derecho. Avda. de la Paz, 137, Logroño, La Rioja. Email: pere.simon@unir.net ORCID ID: https://orcid.org/0000-0003-1722-6498

La publicación es parte del proyecto TED2021-129356B-I00, financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea «NextGenerationEU»/PRTR. Más concretamente, el presente artículo es resultado del proyecto intitulado «Sobre las bases normativas y el impacto real de la utilización de algoritmos predictivos en los ámbitos judicial y penitenciario», con acrónimo Ius-Machina, financiado por el Ministerio de Ciencia e Innovación en la Convocatoria 2021 — «Proyectos de Transición Ecológica y Transición Digital», en el que el autor participa como investigador del grupo de investigación. La fecha de inicio y fin del proyecto es el 01/12/2022 y el 30/11/2024, respectivamente. La cuantía total de financiación asciende a 98.000 euros. IP: Fernando Miró Llinares.

[2] Véanse por todos los interesantes trabajos de Pérez Luño, A. E. (2016). «Nuevo derecho, nuevos derechos». *Anuario de filosofía del derecho*, 32, 15-36; 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. Más concretamente, y al respecto, Pérez Luño ha señalado que «en el horizonte tecnológico del presente,

reciente del pluralismo jurídico, a la apertura jurisdiccional y, también, al creciente impacto de las nuevas tecnologías en el campo del Derecho. Proliferan reconocimientos *ex lege*[3] o por vía jurisprudencial[4] de nuevos derechos individuales y también nuevas garantías vinculadas a estos.

En un entorno mediático en el que crece una tendencia apocalíptica[5] sobre el avance tecnológico, resulta más necesario que nunca recordar que la tecnología es neutra, aunque no los efectos que esta proyecta en su aplicación práctica, motivo por el cual el debate acerca de la introducción de los sistemas de inteligencia artificial[6] en el proceso de toma de decisiones públicas debería girar necesariamente en torno a los usos plausibles, posibles, y a las garantías vinculadas a su empleo[7].

Los tradicionales principios y garantías jurídicas son insuficientes para dar una respuesta adecuada al reto que plantea la realidad algorítmica, tal y como se observa, de un lado, fruto de la imposibilidad de seguir estirando el mandato del art. 18.4 de

---

muchos de los problemas y de las soluciones jurídicas tradicionales aparecen irremediablemente caducos. Esa nueva situación impele al pensamiento jurídico y a la reflexión sobre los derechos a diseñar nuevos instrumentos de análisis y marcos conceptuales prontos para adaptarse a las exigencias de una sociedad en transformación». Pérez Luño, A. E. (2014). «Los derechos humanos ante las nuevas tecnologías», en Pérez Luño, A. E. (ed.), *Nuevas tecnologías y derechos humanos*, Valencia, Tirant lo Blanch, 17.

[3] El Título X de la vigente Ley Orgánica 3/2018, de protección de datos y garantía de derechos digitales es un buen ejemplo. Sobre los nuevos derechos digitales véanse los trabajos de Rallo Lombarte, A. (2019a). «Del derecho a la protección de datos a la garantía de nuevos derechos digitales», en Rallo Lombarte, A. (dir.) *Tratado de protección de datos. Actualizado con la ley orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los derechos digitales*, Valencia, Tirant lo Blanch, 23-52; Rallo Lombarte, A. (2020). «Una nueva generación de derechos digitales». *Revista de Estudios Políticos*, 187, 101-135.

[4] Véanse por todas la STC 290/2000, de 30 de noviembre y la Sentencia del Tribunal de Justicia de la Unión Europea, *Google Spain, S.L. y Google Inc. vs. Agencia Española de Protección de Datos (AEPD) y Mario Costeja González*, ECLI:EU:C:2014:317. Un estudio sobre la citada sentencia relativa al derecho al olvido se encuentra en Simón Castellano, P. (2015). *El reconocimiento del derecho al olvido digital en España y en la UE*, Barcelona, Bosch.

[5] Es la línea adoptada mayoritariamente por los medios de comunicación; sirvan como ejemplo las noticias intituladas «Cómo los algoritmos perpetúan la desigualdad en España», «Un algoritmo impreciso condiciona la libertad de los presos» y «Cómo los algoritmos perpetúan la desigualdad en España», publicados en *La Vanguardia*, disponibles respectivamente en Internet: shorturl.at/fQS05, shorturl.at/aMW89 y shorturl.at/eFHJQ (última consulta el 9 de julio de 2022).

[6] A lo largo de este trabajo, cuando hablamos de inteligencia artificial nos referimos a aquellos sistemas que emulan, imitan o reproducen el pensamiento y obrar humanos —ya sean racionales, humanos, cognitivos o conductistas— e interactúan con el medio, con habilidades tales como la creatividad, la comprensión, la percepción, el lenguaje o el aprendizaje. Para un estudio sobre el contexto y los riesgos de las distintas técnicas relacionadas con los sistemas de inteligencia artificial, véase el capítulo I de la monografía de Hernández Peña, J. C. (2022). *El marco jurídico de la inteligencia artificial. Principios, procedimientos y estructuras de gobernanza*, Cizur Menor, Aranzadi.

[7] Así lo han señalado algunos autores en medios más generalistas. Véase el artículo de opinión de Pere Simón Castellano intitulado «Algoritmos, desigualdad y otros mitos malintencionados», publicado en *El País Cinco Días*, disponible en Internet: shorturl.at/hkORU (última consulta el 9 de julio de 2022).

la Constitución[8] y, del otro, teniendo en cuenta la potencialidad de las herramientas basadas en la llamada inteligencia artificial[9]. Los trabajos previos de los doctores Medina Guerrero[10] y Cotino Hueso[11] advierten de la necesidad de ir más allá de la protección de datos, con un marco jurídico específico para los sistemas de inteligencia artificial[12], tratando de sortear dificultades y obstáculos tales como la preocupante paradoja o falacia de la transparencia[13], sobre la que volveremos más adelante y a la que prestaremos especial atención.

El presente artículo parte de las conclusiones alcanzadas en los estudios citados de la doctrina nacional e internacional[14], y pretende refutar o contrastar las siguientes hipótesis:

Primera.—El debate en torno al marco jurídico del desarrollo de sistemas de inteligencia jurídica artificial y al alcance del derecho fundamental anclado en el art. 18.4 de la CE excede con creces cuestiones meramente semánticas. La necesidad

[8] Es uno de los principales corolarios alcanzados en la jornada «Derechos digitales e inteligencia artificial» organizada por la Fundación Manuel Giménez Abad, coordinada por el catedrático Lorenzo Cotino Hueso, cuyas ponencias y debates están disponibles en abierto: https://www.youtube.com/watch?v=Tiwetu3lC-g (última consulta el 9 de julio de 2022).

[9] Véanse al respecto los trabajos de Bueno de Mata, F. (2020). «Macrodatos, inteligencia artificial y proceso: luces y sombras». *Revista General de Derecho Procesal*, 51; Simón Castellano, P. (2021). *Justicia cautelar e inteligencia artificial: la alternativa a los atávicos heurísticos judiciales*, Barcelona, J. M. Bosch; Simón Castellano, P. (2022). *La prisión algorítmica: Prevención, reinserción social y tutela de derechos fundamentales en el paradigma de los centros penitenciarios inteligentes*, València, Tirant lo Blanch.

[10] 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.

[11] Cotino Hueso, L. (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.

[12] Véanse también las contribuciones al respecto que incorpora el número 100, monográfico sobre inteligencia artificial, de la publicación *El cronista del Estado Social y democrático de Derecho*.

[13] 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.

[14] Los límites espaciales impiden hacer referencias genéricas y profundizar más en qué es un algoritmo y las distintas tipologías de los sistemas de inteligencia artificial, sus posibles usos en determinados ámbitos y los retos que de ella se derivan para distintos campos del conocimiento. Este es un trabajo científico que parte de estudios previos sobre la materia y que pretende avanzar en el conocimiento con una propuesta de taxonomía de las garantías jurídicas relativas al empleo de los sistemas de inteligencia artificial, contribuyendo al debate y en el mejor de los casos, aunque esto es realmente ambicioso, a que se superen los muchos y vacuos estudios que proliferan sobre inteligencia artificial y Derecho, que se limitan a reproducir qué es la inteligencia artificial y sus diferentes tipologías, para incidir en el reto que esto supone para los juristas y finalmente tomar partida con una postura apocalíptica o integrada. Con todo, el autor propone la consulta de algunas obras colectivas introductorias, que reúnen capítulos heterogéneos, también por lo que se refiere a temática —análisis de sectores o impacto en áreas específicas del Derecho—, y que pueden ayudar a tener una visión global sobre la materia. Así, se recomienda la consulta de Huergo Lora, A. J. (Dir.) (2020), *La regulación de los algoritmos*, Cizur Menor, Aranzadi; García Mexía, P. (Dir.) (2022), *Claves de Inteligencia Artificial y Derecho*, Madrid, Wolters Kluwer – La Ley.

de incluir la respuesta del Derecho en el diseño de las soluciones tecnológicas más vanguardistas, en el plano de las garantías y derechos, con una visión sistemática, teniendo presente la dimensión objetiva de los derechos fundamentales y el papel de la dignidad humana y el libre desarrollo de la personalidad, que reconocido en el art. 10.1 de la CE, justo en la antesala del Capítulo segundo del Título primero, implica también importantes consecuencias prácticas.

Segunda.—La transparencia o el derecho de acceder a la información no es la respuesta más adecuada por resultar inidónea, pudiendo incluso producir una falsa sensación de seguridad. En realidad, la transparencia es una pequeña parte de un abanico muy amplio de medidas técnicas y organizativas cuyo fin principal es que el ser humano no sea objeto de decisiones que toman o condicionan las máquinas. Esto es, que como mínimo, los afectados puedan reclamar una «segunda oportunidad"; recurrir, replicar o defenderse con garantías frente a la decisión algorítmica.

Tercera.—Que el debate esté centrado en los usos y en la tecnología, o lo que es lo mismo, en determinar qué tecnología y para qué, permite trazar determinadas líneas rojas, tanto por lo que se refiere a tecnologías invasivas como en lo relativo a las decisiones automatizadas de las que se derivan efectos jurídicos para las personas.

Cuarta.—A pesar de las dificultades y limitaciones para elaborar una teoría general de las garantías jurídicas de los sistemas de inteligencia artificial derivadas de la insuficiencia de los marcos normativos sectoriales y de la necesidad de realizar evaluaciones de impacto *ex ante* en función de qué tecnología (naturaleza y alcance) y para qué (contexto y finalidades), resulta conveniente realizar un esfuerzo de dogmática aplicada que culmine con una propuesta con fines utilitarios o pragmáticos. La transparencia y la explicabilidad gozan de autonomía conceptual y, aunque pueden perseguir unos mismos fines —una IA confiable y garantizar la vía del recurso frente a las decisiones algorítmicas—, constituyen garantías relativas que pueden o deben compaginarse, en función de cada caso concreto, con garantías reforzadas tales como la seguridad —cuya propiedad más significativa es la trazabilidad—. La participación humana se ha apuntado en muchas ocasiones como una panacea que, por el contrario, también debe ser relativizada. Finalmente, encontramos también las garantías institucionales —marco sancionador y organismo de control, europeo y nacional—.

Con tal fin, el trabajo parte de lo general para terminar en lo concreto, esto es, de la dimensión objetiva de los derechos fundamentales hasta las garantías institucionales de los sistemas de inteligencia artificial.

II. MÁS ALLÁ DE CUESTIONES SEMÁNTICAS: LA LEY LIMITARÁ EL USO DE LA INFORMÁTICA O LA DIMENSIÓN OBJETIVA DE LOS DERECHOS FUNDAMENTALES.

El legislador constitucional español estuvo especialmente despierto al incorporar en el art. 18.4 de la CE un mandato al legislador, en forma de limitación genérica

© UNED. *Revista de Derecho Político*
N.º 117, mayo-agosto, 2023, págs. 153-196

de la informática, con objeto, al menos inicialmente, de proteger el honor y la intimidad personal y familiar. Tal mandato ha sufrido hasta día de hoy una importante evolución interpretativa, hasta el punto de convertirse en un derecho fundamental, autónomo e independiente de cualquier otro.

Más concretamente, el mandato constitucional al poder legislativo se ha traducido en una regulación estatal que se encuentra fuertemente condicionada por la normativa europea, hasta tal punto que algunos autores como Rallo Lombarte han llegado a señalar de forma contundente que se ha producido «la abducción de un derecho constitucional convertido en un derecho exclusivamente europeo»[15].

La consagración de un derecho europeo a la protección de datos, que incorpora amplias y genéricas definiciones de dato personal y de tratamiento de datos, ha conllevado en la práctica que nuestra Ley Orgánica de Protección de datos[16] actúe a modo de complemento de la normativa europea[17], que en forma de Reglamento goza de aplicabilidad directa, y que el contenido total tutelado del citado derecho fundamental dependa, en exclusiva, en términos empleados por Medina Guerrero, de «lo que el RGPD[18] dice que es»[19].

El modelo de tutela *sui generis* establecido por la normativa europea no encuentra parangón con otros derechos fundamentales reconocidos *ex constitutione*. Resulta imposible realizar una analogía incluso con otros derechos de la personalidad, cuya tutela se vehicula a través de procesos civiles por responsabilidad extracontractual y de indemnización por daños ilegítimos[20]. En cambio, y al margen de la existencia de un derecho de indemnización en el ámbito de la protección de datos que nunca ha tenido un impacto tangible ni un desarrollo real[21], se ha construido un derecho

---

[15] Rallo Lombarte, A. (2019b). «El nuevo derecho de protección de datos». *Revista Española de Derecho Constitucional*, 116, 45-74.

[16] Nos referimos a la actual Ley Orgánica 3/2018, de 5 de diciembre (en adelante, LOPDGDD).

[17] Lucas Murillo De La Cueva, P. (2020). «Encuesta sobre la protección de datos personales». *Teoría y Realidad Constitucional*, 46, 40.

[18] Acrónimo del Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (en adelante, RGPD).

[19] Medina Guerrero, M. (2022), cit., 144.

[20] La extinta Directiva Directiva 95/46/CE, aprobada en 1995, en su artículo 23.1, ya establecía que «Los Estados miembros dispondrán que toda persona que sufra un perjuicio como consecuencia de un tratamiento ilícito o de una acción incompatible con las disposiciones nacionales adoptadas en aplicación de la presente Directiva, tenga derecho a obtener del responsable del tratamiento la reparación del perjuicio sufrido».

[21] El art. 23.1 de la Directiva 95/46/CE tuvo su reflejo en el art. 19 de la Ley Orgánica 15/1999, de 13 de diciembre, y un desarrollo más decidido en el art. 82 del RGPD, si bien la LOPDGDD no contiene ninguna mención al derecho a obtener una indemnización derivada de los daños y perjuicios materiales o inmateriales como consecuencia de una infracción de la normativa sobre protección de datos. Sea como fuere, la normativa siempre ha sido ambigua en el sentido de no resolver cuales son los perjuicios a los que se extiende la indemnización, ni establecer cuál es la forma en la que se tienen que valorar los daños ni las excepciones en la responsabilidad legal; precisamente, por eso, ha recibido notorias

europeo de carácter administrativo, cuya tutela se vehicula a través de la existencia de autoridades de control fuertes, con funciones de inspección y sancionadoras, e incluso de interpretación extensiva de las obligaciones establecidas *ex lege*[22]. Unos poderes de supervisión y tutela administrativa que irradia y que en muchas ocasiones ha supuesto en la práctica una protección indirecta de otros derechos de la personalidad, llegando la AEPD a entrar a resolver asuntos que no son propiamente una cuestión de protección de datos, sino de temas afines, lo que ha llevado a algunos autores a hablar de «poderes salvajes de las autoridades de control»[23] o de tutela desbocada.

Sea como fuere, en este estado de cosas, ¿es la vía de la protección de datos la mejor fórmula para encauzar y dar respuesta a los retos que plantea el desarrollo tecnológico de sistemas de inteligencia artificial?

II.1. El derecho a la protección de datos: un enfoque insuficiente

Se ha pretendido dar respuesta a los retos planteados por las tecnologías disruptivas y la inteligencia artificial a través de la normativa de protección de datos En buena medida, se trata de una pretensión loable ante la falta de instrumentos o mecanismos legales que permitan dar una respuesta uniforme, así que no es de extrañar que algunas autoras como Ni Loideain se hayan referido al RGPD como un «refugio» frente al *Internet of Things*, el empleo de datos masivos —*Big Data*— o las soluciones basadas en inteligencia artificial[24]. Sin embargo, se trata de un marco legal atomizado e insuficiente. Como indica Sarrión Esteve, en realidad estamos ante «un nuevo paradigma en la protección de los derechos fundamentales», por lo que resulta necesaria «una reorientación {*o reinterpretación de los principios jurídicos tradicionales*} a consecuencia de la naturaleza de las tecnologías de las que hablamos»[25].

El RGPD ofrece una serie de principios, normas y derechos que pueden tener un impacto escaso, parcial o restringido sobre los proyectos de desarrollo tecnológico avanzado. Los sistemas de inteligencia artificial implican siempre tratamiento de

críticas de la doctrina. Véase al respecto los trabajos de Grimalt Servera, P. (1999). *La responsabilidad civil en el tratamiento automatizado de datos personales*, Colección Estudios de Derecho Privado núm. 8, Granada, Comares, pp. 88 y 89; Herrán Ortiz, A. I. (2003), *El derecho a la protección de datos personales en la sociedad de la información*, Bilbao, Cuadernos Deusto de Derechos Humanos núm. 26, p. 73.

[22] Véanse las muchas guías para responsables y encargados de tratamiento, que inciden desde las obligaciones de informar hasta cuestiones tan enconadas como cómo realizar un análisis de riesgo o una evaluación de impacto en protección de datos. Pueden consultarse en la web de la AEPD: https://www.aepd.es/es (última consulta el 9 de julio de 2022).

[23] Simón Castellano, P. y Dorado Ferrer, J. (2022). «Límites y garantías constitucionales frente a la identificación biométrica». *Revista de Internet, Derecho y Política (IDP)*, 35, 1-13, 11.

[24] Véase Ni Loideain, N. (2018). «A Port in the Data-Sharing Storm: The GDPR and the Internet of Things». *King's College London Law School Research Paper*, 2018-27.

[25] Sarrión Esteve, J. (2020). «El derecho constitucional en la era de la inteligencia artificial, los robots y los drones», en Pérez Miras, A. *et. al* (Dirs.), *Setenta años de Constitución Italiana y cuarenta años de Constitución*, Vol. 5, Madrid, Centro de Estudios Políticos y Constitucionales, 321-334, p. 328.

datos aunque sus principios tienen una aplicación limitada como muy bien apunta Cotino Hueso[26]: la limitación de la finalidad del tratamiento (art. 5.1.b RGPD) en relación con el uso y descubrimiento de nuevas correlaciones; la limitación del plazo de conservación (art. 5.1.e RGPD) puesto que es natural que se encadenen unas y otras investigaciones a partir de los hallazgos de la anterior; el binomio derecho deber de transparencia e información al interesado (arts. 5.1.a y 14.5.b RGPD) frente a usos insospechados o desconocidos en el momento de la obtención del consentimiento; la minimización de datos (art. 5.1.c RGPD) porque va en contra de muchos sistemas que se basan precisamente en mezclar grandes cantidades de información —datos personales y otros datos que no tienen esa consideración—.

Además, ni el consentimiento informado y explícito ni el interés legítimo resultan bases de legitimación aplicables en muchos entornos de desarrollo de sistemas de inteligencia artificial. Por ello, autores como Mantelero se refieren a la necesidad de «ir más allá de los datos [*personales*]»[27].El derecho a la protección de datos, en definitiva, construido como un derecho europeo y de tutela administrativa, se ha configurado como un derecho subjetivo, individual, resultando en muchas ocasiones un instrumento incapaz o no pensado para dar respuesta a las consecuencias sociales que sobrepasan el ámbito del interesado. En la misma dirección, y de forma preclara, Cotino Hueso se refiere a que «el derecho individual y subjetivo de protección de datos es insuficiente frente al poder efectivo que implica el enorme poder real que confiere el manejo masivo de datos»[28].

Se trata de una idea a la que se ha referido el supervisor europeo de forma muy gráfica cuando nos dice que «se van dejando caer migas digitales a cada minuto, que se combinan para clasificar a las personas físicas en tiempo real y para crear perfiles múltiples y, en ocasiones, contradictorios»[29]. El análisis micro o la respuesta a nivel individual es absolutamente insuficiente frente al desarrollo de soluciones disruptivas que implican técnicas basadas en inteligencia artificial[30]. El uso y empleo de los datos, personales o no, afecta al derecho de protección de datos personales como

[26] Cotino Hueso, L. (2022a), cit., 85.

[27] El autor citado comprende que el enfoque correcto ante los sistemas de inteligencia artificial debe partir de un modelo de responsabilidad proactiva con análisis específico del potencial impacto de la tecnología sobre los derechos de las personas —no discriminación, igualdad, posibilidad de recurso, etc.—, con una lectura sistemática que trasciende la protección de datos, y que incluye cuestiones derivadas del impacto social y ético de las herramientas más vanguardistas. Véase Mantelero, A. (2022). *Beyond Data. Human Rights, Ethical and Social Impact Assessment in AI,* La Haya, Springer, pp. 1-91 y 185-197.

[28] Cotino Hueso, L. (2022a), cit., 80.

[29] Supervisor Europeo de Protección de Datos (en adelante, SEPD), *Dictamen 4/2015. Hacia una nueva ética digital. Datos, dignidad y tecnología*, de 11 de septiembre de 2015, p. 15.

[30] En la misma dirección, de nuevo, resultan de referencia obligada los trabajos de Mantelero, A. (2018). *El big data en el marco del Reglamento General de Protección de Datos*, Barcelona, UOC, 1-46; Mantelero, A. (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.

autodeterminación informativa, pero va más allá de la dimensión individual que este confiere y adquiere una dimensión colectiva, que como decíamos anteriormente no obtiene tutela suficiente mediante el actual marco normativo de protección de datos.

El legislador europeo parece ser consciente de las limitaciones de la normativa de protección de datos en este ámbito y por ello ha avanzado con una propuesta de Reglamento específico con el fin de establecer normas armonizadas en materia de inteligencia artificial[31] (en adelante, LIA, acrónimo de Ley de Inteligencia Artificial, terminología que la propia propuesta incluye en su versión en castellano). Tal preocupación parece también encontrarse en los motivos que subyacen a la primera regulación específica de la inteligencia artificial en España; nos referimos a la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, que contiene la primera regulación positiva sobre el empleo de la inteligencia artificial por las administraciones públicas y las empresas en nuestro país, con una estructura programática y propositiva, que diseña líneas de actuación de las administraciones públicas con el objetivo de favorecer, promover y priorizar determinadas políticas y prácticas relacionadas con el uso de «algoritmos involucrados en la toma de decisiones».

En ambos casos, el legislador —europeo y nacional— abraza posturas cuyo enfoque nuclear es mucho más amplio que la mera protección de datos personales, haciendo referencia a conceptos que incluyen la transparencia y la rendición de cuentas, la minimización de sesgos para abordar un potencial impacto discriminatorio y mecanismos y herramientas para un diseño racional de los sistemas algorítmicos[32].

Con todo, la Ley 15/2022, de 12 de julio, constituye la primera norma que en nuestro ordenamiento señala cómo deberán diseñar las Administraciones Públicas los algoritmos utilizados en esa toma de decisiones, y cuya principal razón de ser es la inexistencia de un marco jurídico estable que permita dar respuestas uniformes a los entornos de desarrollo de soluciones basadas en inteligencia artificial.

II.2. ¿Un estándar de transparencia?

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

[31] Propuesta de Reglamento del Parlamento Europeo y del Consejo, de fecha 21 de abril de 2021, disponible en Internet: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex:52021PC0206 (última consulta el 9 de julio de 2022).

[32] En Europa, esta visión la encontramos también en documentos de trabajo de organismos como el consorcio *AI Ethics Impact Group* (en adelante, AIEI Group), liderado por la Asociación VDE de Tecnologías Eléctricas, Electrónicas y de la Información y *Bertelsmann Stiftung*. Más concretamente, el AIEI Group propone el establecimiento de seis valores clave que sirvan de baremo y que son la transparencia, la responsabilidad, la privacidad, la justicia, la confiabilidad y la sostenibilidad ambiental. Véase AIEI Group (2020). «AI Ethics Impact Group: From Principles to Practice – An interdisciplinary framework to operationalise AI ethics», disponible en Internet: https://www.ai-ethics-impact.org/en (última consulta el 9 de julio de 2022).

vehiculado y centrado en una suerte de derecho a conocer el código algorítmico y, en el mejor de los casos, también su funcionamiento[33]. Nos referimos al acceso al código fuente de la solución tecnológica, que se ha planteado como solución por ejemplo en el caso de Eric Loomis[34] y también en el caso BOSCO[35]. En ambos supuestos la

[33] Especialmente por parte de la doctrina administrativista. Véanse los trabajos de Cerrillo i Martínez, A. (2020). «La transparencia de los algoritmos que utilizan las administraciones públicas». *Anuario de Transparencia Local,* 3, 41-78; 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. Para un enfoque vinculado a la protección de datos y conectado con el derecho de acceso del art. 15 del RGPD, véase Medina Guerrero, M. (2022), cit., pp. 162 y ss.

[34] Nos referimos al caso *Wisconsin v. Loomis*, en el que el Tribunal Supremo de ese estado resuelve el caso de un particular, Eric Loomis, que había sido acusado de cinco delitos por participación en un tiroteo efectuado desde un vehículo. El acusado negó su participación en el tiroteo, aunque admitió haber conducido el vehículo con posterioridad a los hechos, y llegó a un acuerdo con la Fiscalía para aceptar los dos cargos y rechazar el resto de las acusaciones. El juez de instancia aceptó la conformidad y al concretar la pena tuvo en cuenta, entre otras cosas, el contenido de un *presentence investigation report* basado en el resultado de la evaluación de Eric Loomis mediante la herramienta de inteligencia jurídica artificial para la valoración del riesgo COMPAS, que valora el riesgo de reincidencia y las necesidades criminógenas del sujeto. La citada herramienta arrojaba en relación con Eric Loomis un riesgo alto de reincidencia, en general, y de reincidencia violenta, en particular. El juez condenó a Loomis a una pena de seis años de prisión y otros 5 de supervisión post penitenciaria, razonando que esta pena es el resultado de la valoración de distintos factores entre los que se encuentra el riesgo extremadamente alto de reincidir que le atribuye la herramienta COMPAS, excluyendo la posibilidad de suspender la condena. Como reacción, Eric Loomis interpuso recurso ante el Tribunal Supremo del estado de Wisconsin, alegando que el uso de la herramienta de inteligencia jurídica artificial para la valoración del riesgo había vulnerado su derecho a un juicio justo

Por basarse en información poco fiable o precisa. Ninguno de los motivos del recurso prosperó; para apoyar esta posición, el Tribunal señaló que la exactitud de los instrumentos utilizados y la capacidad de los jueces para entender su posible mal funcionamiento eran suficientes para asegurar los derechos de los acusados, amparando la opacidad del algoritmo en base a los derechos de propiedad intelectual sobre la herramienta. Sobre este asunto véanse los trabajos de Martínez Garay, L. (2018). «Peligrosidad, algoritmos y due process: El caso State vs. Loomis». *Revista de Derecho Penal y Criminología*, 20, 485-502, en especial p. 492; De Miguel Beriain, I. (2018). «Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the Wisconsin v. Loomis ruling». *Law, Probability and Risk*, vol. 17, 1, 45-53; Simón Castellano, P. (2022), cit., pp. 112-114.

[35] Se trata de un programa informático empleado por las compañías eléctricas, que ha sido desarrollado por el Gobierno con el fin de evaluar las solicitudes de aquellos que pretenden acogerse al bono social que genera un descuento en la factura de la luz. Puede accederse al documento de comprobación de las condiciones de otorgamiento del Bono Social en abierto: https://civio.app.box.com/s/j4n7nk3e3wmxq5lk5d3otckif1jv3z50 (última consulta el 9 de julio de 2022) y a la Sentencia núm. 143/2021, del Juzgado Central de lo Contencioso Administrativo núm. 8, en el procedimiento ordinario 18/2019, de 30 de diciembre, que entiende que la liberación del código, además de contravenir la propiedad intelectual, afectaría tanto a la seguridad pública como a la defensa nacional, disponible en https://civio.app.box.com/s/ok4x3w0l12iougk4cz6mt5ec73vdafb9 (última consulta el 9 de julio de 2022).

negativa o el rechazo a ofrecer el código se fundamentó en base a la existencia de derechos de propiedad industrial sobre la herramienta tecnológica.

Otra analogía posible en perspectiva comparada la encontramos en Brasil, país en el que se permite para el sistema electrónico electoral el acceso al código fuente, como garantía de su funcionamiento neutral, lo que resulta posible a diferencia de los dos ejemplos anteriores en la medida que el Tribunal Supremo Eleitoral es el titular del derecho de propiedad sobre el *software* desarrollado bajo la supervisión de la Administración electoral[36].

Sin embargo, la transparencia o el acceso íntegro al código fuente stricto sensu no es un fin en si mismo y no actúa a modo de panacea frente a los posibles usos no deseados de los algoritmos. Lo expresa claramente Medina Guerrero cuando advierte que «el objetivo de lograr la rendición de cuentas de las decisiones algorítmicas imponiendo su explicación a los afectados no es sino 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»[37].

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[38]— sin posibilidad de 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

[36] Véase Presno Linera, M. A. (2016). «Premisas para la introducción del voto electrónico en la legislación electoral española». *Revista de Estudios Políticos,* 173, 277-304, p. 297.

[37] Medina Guerrero, M. (2022), cit., p. 169.

[38] 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.

información y de la metodología de los análisis de riesgos propios de los sistemas de cumplimiento normativo[39].

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[40].

II.3. La dignidad humana y el libre desarrollo de la personalidad

De la insuficiencia de los enfoques previos deriva la necesidad de acudir al mandato del art. 10.1 de la CE por lo que se refiere a la dignidad humana y el libre desarrollo de la personalidad[41], y a la llamada dimensión objetiva de los derechos,

[39] 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.

[40] 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 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).

[41] En una dirección parecida, Presno Linera indica que «el recurso a sistemas de IA por parte de los poderes públicos o de entidades que actúen en su lugar debe hacerse sin que quede afectado ese mínimo invulnerable que garantiza el principio de la dignidad humana y, en la medida de lo posible, promoviendo el libre desarrollo de la personalidad». Presno Linera, M. A. (2022c). *Derechos fundamentales e Inteligencia Artificial*, Madrid, Marcial Pons, Fundación Manuel Giménez Abad, 104.

que compartimos con otros autores debería actuar como «catalizador jurídico constitucional para el tratamiento de la inteligencia artificial»[42].

La existencia de la dimensión objetiva de los derechos fundamentales implica aceptar que estos tienen significado propio en su alcance y contenido, pero también cobran sentido en su relación global con el sistema de derechos establecido por la Constitución, lo que a vueltas afecta a la concreción de ese ámbito y contenido[43]. La dimensión objetiva actúa como límite de intervención empero también como apoyo de pretensiones de prestación frente al Estado, como deber positivo de protección de estos cuando la libertad garantizada o tutela es puesta en peligro por parte de terceros.

Así las cosas, los poderes públicos tienen en su ámbito de actuación un deber de realización que implica en la mayoría de las ocasiones un deber de actividad positiva por parte del Estado de muy diversa índole. En el ámbito de la inteligencia artificial y las tecnologías disruptivas, bien podría promover o exigir el establecimiento de un marco jurídico armonizado que de forma integral dé respuesta a retos jurídicos que no pueden abordarse con garantías si se hace de forma limitada, sectorial o atomizada. Podría entonces actuar la dimensión objetiva como fundamento para exigir que el desarrollo tecnológico basado en herramientas de inteligencia artificial se realice con respeto a la dignidad humana, al libre desarrollo de la personalidad y a ciertos principios éticos que tienen pleno encaje y se ven reflejados en nuestras actuales cartas de derechos fundamentales —consensos relativos a la necesaria ausencia de sesgos y prohibición de discriminación, libertad y seguridad de los individuos, principio de legalidad, derecho de defensa, etc.—, aunque con una perspectiva que trasciende a los riesgos y amenazas concretos o a los derechos individuales, y que se centra en la posible afectación a grandes colectivos o a la propia humanidad en su conjunto[44].

Sea como fuere, y siguiendo el planteamiento ofrecido por Cotino Hueso, la dimensión objetiva de los derechos es un recurso jurídico y dogmático especialmente útil ante las necesidades que se derivan debido al impacto de la inteligencia artificial[45]. Este es también uno de los principales corolarios alcanzados en los debates que la doctrina ha mantenido al respecto de esta cuestión en los últimos tiempos[46].

Otra fortaleza nada desdeñable de este enfoque es que la dimensión objetiva se traduce, también, en obligaciones de eficacia de los derechos fundamentales en las relaciones entre particulares. Especialmente si tenemos en cuenta que el desarrollo

[42] Cotino Hueso, L. (2022a), cit., 74.

[43] Véase al respecto la delimitación conceptual ofrecida por De Otto y Pardo en Martín Retortillo, L. y De Otto y Pardo, I. (1988). *Derechos fundamentales y Constitución*, Madrid, Civitas, 163 y ss.

[44] Sobre el libre desarrollo de la personalidad y su proyección sobre concretos derechos fundamentales véase Presno Linera, M. A. (2022b). *Libre desarrollo de la personalidad y derechos fundamentales,* Madrid, Marcial Pons.

[45] Cotino Hueso, L. (2022a), cit.

[46] Véase la jornada «Derechos digitales e inteligencia artificial» organizada por la Fundación Manuel Giménez Abad, cfr. nota 7 del presente trabajo, y más concretamente las ponencias de los doctores Rallo Lombarte, Cotino Hueso, Teruel Lozano y Simón Castellano.

de sistemas de inteligencia artificial, hasta la fecha, ha estado en manos del sector privado o de las grandes tecnológicas[47], que aparece hoy como la gran amenaza para determinados bienes jurídicos objeto de interés y tutela constitucional[48]. En este ámbito, las obligaciones derivadas de los derechos para el sector privado deben articularse a través de la regulación o bien como mandatos de interpretación de las normas entre particulares. Con base en la dimensión objetiva de los derechos se legitima jurídicamente la creación de entidades administrativas de tutela e interpretación de derechos fundamentales, como las ya existentes autoridades de control en materia de protección de datos[49] o las futuras autoridades de inteligencia artificial como la que prevé la LIA, sobre la que volveremos en el epígrafe IV.5 del presente trabajo, al que nos remitimos.

III. HACÍA UN DEBATE CENTRADO EN LOS USOS

Identificar un conjunto común de principios éticos y sociales compartidos, que luego se traducirán en garantías y derechos específicos, que se tomen como referencia para evaluar las posibles consecuencias de las aplicaciones basadas en inteligencia artificial, es un ejercicio complejo y para nada pacífico, si tenemos en cuenta las múltiples tecnologías posibles que se agrupan bajo la voz de inteligencia artificial, los valores que inspiran los modelos lógicos —racionales o humanos; cognitivos o conductistas— que están detrás del funcionamiento de los algoritmos y el hecho que el desarrollo tecnológico no es estático y que puede variar de un contexto cultural a otro[50].

---

[47] Hemos analizado críticamente *ut supra* los problemas en relación con la propiedad intelectual derivados de la creación, mantenimiento e implementación de los algoritmos por parte de empresas privadas, como en el caso COMPAS o BOSCO, lo que luego en la práctica se traduce en una opacidad insufrible que produce un menoscabo sobre las facultades de defensa de los individuos.

[48] Un fenómeno parcialmente nuevo puesto que el derecho a la intimidad, el derecho no ser molestado o *The Right to Privacy*, siguiendo la terminología propuesta por Warren y Brandeis a finales del siglo XIX, se configuró inicialmente como un derecho de los individuos a proteger su vida privada frente a la tentación por parte de los poderes del Estado de inmiscuirse en esa esfera reservada. Hoy en día, en cambio, las principales amenazas para la privacidad de los individuos proceden de las grandes tecnológicas, a las que los usuarios en ocasiones ceden voluntariamente sus datos bajo la creencia de que se ofrecen aplicaciones o servicios de forma gratuita. Véase Warren, S. D. y Brandeis, L. D. (1890). «The Right to Privacy». *Harvard Law Review*, 4, 5, 193-220.

[49] Aunque como se ha indicado anteriormente resulta difícil delimitar los límites de intervención por parte de estas, cuanto menos cuando interpretan y ponderan otros derechos fundamentales en conflicto, lo que ha sido objeto de crítica doctrinal. Véase Simón Castellano, P. y Dorado Ferrer, J. (2022), cit., 11.

[50] La misma noción de privacidad varía en el debate transatlántico, lo que en la práctica ha llevado al Tribunal de Justicia de la Unión Europea a declarar inválidas las dos decisiones de adecuación de la Comisión Europea conocidas bajo las voces de puerto seguro —*Safe Harbor*— y escudo de privacidad —*Privacy Shield*—, puesto que el nivel de protección garantizado en el país tercero, en este caso en EE. UU., no es equivalente ni ofrece garantías adecuadas. Véanse las SSTJUE de 6 de octubre de 2015,

A pesar de ello, la doctrina ha identificado en relación con determinadas tecnologías, como es el caso del *Big Data*, un posible marco común de referencia con base en los valores reconocidos por las cartas internacionales de derechos humanos[51]. Se trata en cualquier caso de una indicación abstracta y limitada, y por ende siempre resultará necesario proceder a una ponderación *ad hoc* para el caso concreto, teniendo en cuenta la naturaleza específica y el contexto de aplicación de la tecnología en cuestión, así como la proporcionalidad —entre otros, idoneidad y necesidad[52]— en relación con la finalidad perseguida, los medios empleados y el impacto previsto para los derechos de las personas afectadas o sobre las cuales el algoritmo proyecta sus propuestas y decisiones.

Ahora bien, ¿quién debe realizar esa aproximación o evaluación? Se impone en nuestro modelo una perspectiva *ex ante* que se refuerza en el contexto global de los *compliance* y de las normas internacionales o estándares voluntarios como la ISO, propios de sistemas de seguridad de la información, en los que se inspira la LIA.

III.1. El enfoque de riesgo propuesto por la UE.

El objetivo del nuevo instrumento (LIA) 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

asunto C-362/14 (Schrems), ECLI:EU:C:2015:650; de 16 de julio de 2020, Case C-331/18 (Schrems II), ECLI:EU:C:2020:559.

[51] Wright, D. (2011). «A framework for the ethical impact assessment of information technology». *Ethics Inf Technol,* 13, 3, 199-226.

[52] Un excesivo rigor o formalismo con el principio de necesidad podría actuar como un auténtico lastre para la innovación tecnológica. La interpretación reciente de la AEPD en relación con el empleo de sistemas biométricos —huella dactilar, reconocimiento facial, etc.— para el registro de jornada —obligación legal— en cierto modo condena a las empresas y organismos públicos a alternativas que no son para nada razonables. Evidentemente podemos encontrar medios menos intrusivos —el papel, sin ir más lejos— que no requieren biometría para el registro de jornada de empleados y trabajadores, y sin embargo son «alternativas» menos seguras que permiten vaciar de contenido una obligación legal cuya principal finalidad no es otra que proteger los derechos de los trabajadores. Véase por todas la Resolución de la AEPD en el PS/00218/2021. Otro ejemplo lo encontramos en el informe de la AEPD en materia de reconocimiento facial aplicado a las universidades online (*e-proctoring*), que rechazó el uso de esta tecnología para garantizar que los estudiantes no copiaran en sus los exámenes, esgrimiendo que la medida no superaba el test de necesidad, por existir medios o vías menos gravosas. Lo cierto es que, en este caso, teniendo en cuenta el escenario creado por la pandemia por COVID-19, es difícil incluso determinar cuáles son esas alternativas menos gravosas a las que se refiere la AEPD, que permitían a los estudiantes examinarse. Véase el Informe sobre reconocimiento facial en los exámenes de la AEPD, N/REF: 0036/2020, p. 16.

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.

Al mismo tiempo, se han elaborado nuevas reglas sobre maquinaria que utiliza inteligencia artificial, para dar mayor confianza a los usuarios tanto para el empleo como para la generación de nuevos productos a través de ella. La citada propuesta se construye, al menos parcialmente, sobre el modelo de la legislación preexistente relativa a la seguridad de los productos, si bien sus criterios y reglas pueden proyectarse hacía terceros estados, como el efecto expansivo que, ya ha ocurrido anteriormente, por ejemplo, con el RGPD[53].

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, 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*[54] (en

[53] Véase al respecto el trabajo de Rustad, M. L. y Koenig, T. H. (2019). «Towards a global data privacy standard». *Florida Law Review*, 71, 365-454.

[54] 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

adelante, DSA), 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.

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[55]. 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

---

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

[55] Véase 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).

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 cao 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.

La doctrina nacional[56] e internacional[57] 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 LIA 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[58].

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 LIA. 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[59], se señala que el enfoque

---

[56] 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,* disponible en Internet: https://diariolaley.laleynext.es/Content/DocumentoRelacionado.aspx?params=H4sIAAAAAAAAEAMtMSbF1CTEAAmMjS0NTS7Wy1KLizPw827DM9NS8klS15Jz UxCKXxJJU58Sc1LyUxCLbkKLSVABz2GlLNwAAAA==WKE#I21(última consulta el 9 de julio de 2022).

[57] 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).

[58] 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).

[59] Véase Cotino Hueso, L. *et al.* (2021), cit.

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 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 delimitación de los sistemas de alto riesgo y, también, de un mecanismo de actualización *ad hoc*[60].

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[61]. 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[62].

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[63]. 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[64]; 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

---

[60] Ibídem.

[61] Véase Miguel Asensio, P. A. (2021). «Propuesta de Reglamento sobre inteligencia artificial». *La Ley Unión Europea*, núm. 92.

[62] Posterior a la redacción del presente trabajo, con fecha de 28 de septiembre de 2022, se ha publicado la 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), cuyo texto en caso de ser aprobado cubriría la laguna que se señala en el cuerpo del trabajo además de complementar en buena medida la LIA. La propuesta de Directiva citada se encuentra disponible en Internet: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52022PC0496 (última consulta el 16 de diciembre de 2022).

[63] Smuha, N. A. *et al.* (2021), cit.

[64] 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.

y la democracia, y por lo tanto no garantiza una inteligencia artificial basada en los principios de legalidad, ética y robustez[65]. 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 de la LIA un derecho explícito de reparación para las personas que 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.

III.2. Líneas rojas: decisiones automatizadas, fines y tecnologías prohibidas

El artículo 5 de la LIA 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[66]; 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

[65] Smuha, N. A. *et al.* (2021), cit.

[66] Lo que significa una prohibición de buena parte de las técnicas del llamado *predictive policing*, cuya finalidad en principio no es otra que la prevención e investigación de la delincuencia. Las técnicas y tecnologías que conforman el *predictive policing* son diversas, pero algunas incluyen sistemas de identificación biométrica y se emplean bases de datos de delitos graves que se mezclan con datos relativos a infracciones leves. La doctrina ha criticado que esta técnica acaba generando más atención policial a los barrios donde tales alteraciones se producen lo que, a su vez, se traduce en mayor número de identificaciones y detenciones, generando así una retroalimentación donde la geografía opera como un sustitutivo de la raza. Al respecto véanse Ferguson, A. G. (2017). *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, Nueva York, New York University Press; Miró Llinares, F. (2020). «Predictive policing: Utopia or dystopia? On attitudes towards the use of big data algorithms for law enforcement». *Revista de Internet, Derecho y Política (IDP)*, 30.

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 considera 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[67] 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

[67] 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.

impone ciertos requisitos de transparencia, exigiendo que los sistemas de inteligencia artificial que interactúen con humanos deban informar del hecho que estos están hablando 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 LIA, 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[68].

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. La necesidad de realizar una autoevaluación previa nos recuerda y mucho la previsión de la DSA, comentada *ut supra*, de «evaluación de riesgos sistémicos». Bien podría interpretarse que esos requisitos y obligaciones reforzadas de los sistemas de riesgo alto —tanto para desarrolladores como para importadores y proveedores— son también «sistémicos», por tratar de mitigar los niveles de riesgo residual que se le atribuyen —de entrada, con una categorización basada en el tipo de tecnología y uso previsto— y que, por ello, incorporan medidas que pueden vincularse con voces tan dispares como la seguridad, privacidad, transparencia, responsabilidad, justicia, confiabilidad y sostenibilidad ambiental.

III.3. La responsabilidad proactiva y las (auto)evaluaciones de impacto

Como hemos observado, proliferan textos internacionales, aunque también otros de ámbito nacional o incluso regional[69], que incorporan alertas, recomendaciones o

[68] La propuesta de Reglamento también ha sido objeto de estudio y análisis crítico en Simón Castellano, P. (2022), cit., 143-153.

[69] Sirva como ejemplo el caso de Cataluña, que puso en marcha en febrero de 2020 la *Estratègia d'Intel·ligència Artificial de Catalunya* con el fin de fortalecer el ecosistema catalán de inteligencia artificial, con un eje específico de «ética y sociedad». Para desarrollar el citado eje, se creó el *Observatori d'Ètica en Intel·ligència Artificial de Catalunya* (en adelante, OEIAC) que publica informes sobre la materia. Véanse

directrices que, basadas en principios éticos, tratan de modular e informar el desarrollo de soluciones o herramientas basadas en inteligencia artificial. Además, la mayoría de esos textos resultan complejos de entender porque están escritos por académicos o técnicos[70], dificultando su comprensión también para las empresas —potenciales desarrolladoras y proveedoras—, administraciones, organismos profesionales, etc., lo que a la postre retrasa su adopción y aplicación práctica[71].

Sin embargo, y mientras CEN-CENELEC[72] e ISO[73] aún están en una primera fase de desarrollo y discusión de los estándares técnicos para la inteligencia artificial, se han publicado algunas alternativas que más allá de una definición de recomendaciones éticas, tratan de establecer modelos basados en cuestionarios y formularios que permitan realizar una autoevaluación de impacto de la herramienta tecnológica.

Una referencia indiscutible es el método de evaluación de impacto en derechos fundamentales y algoritmos (en adelante, por sus siglas en inglés, FRAIA) que el gobierno de Holanda ha creado como herramienta útil para la evaluación en caso de organizaciones gubernamentales. La FRAIA incorpora un exhaustivo cuestionario que tiene en cuenta el diseño y la propiedad sobre la herramienta, la posibilidad de delegar su desarrollo o mantenimiento, los ajustes y mejoras, así como el empleo del algoritmo en la práctica[74]. Los resultados arrojan las debilidades y los principales focos de riesgo del algoritmo de una forma estructurada para que los responsables en cuestión implementen en consonancia controles que permitan mitigarlos.

Otra iniciativa interesante para simplificar el proceso de autoevaluación es la que ha ofrecido el Gobierno de Canadá, pionero en la materia, con el estándar *Algorithmic Impact Assessment Tool*, que permite determinar el nivel de impacto de un sistema que

---

algunos de los resultados en OEIAC. (2021). «Inteligencia artificial, ética y sociedad: Una mirada y discusión a través de la literatura especializada y de opiniones expertas», disponible en Internet: https://www.udg.edu/ca/Portals/57/OContent_Docs/Informe_OEIAC_2021_cast-4.pdf (última consulta el 9 de julio de 2022); OEIAC (2021). «El Model PIO (Principis, Indicadors i Observables): Una proposta d'autoavaluació organitzativa sobre l'ús ètic de dades i sistemes d'intel·ligència artificial», disponible en Internet: https://www.udg.edu/ca/Portals/57/OContent_Docs/modelPIO_v6.pdf (última consulta el 9 de julio de 2022).

[70] Diagnóstico y crítica efectuada previamente en el trabajo de Jobin, A. *et al.* (2019. «The global landscape of AI ethics guidelines». *Nat Mach Intell,* 1, 389–399.

[71] En esa misma dirección, y en favor de un visión coordinada y más sofisticada, véase Morley, J. *et al.* (2020). «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.

[72] Se trata del Comité Europeo de Normalización Electrotécnica que, creado en 1973, es responsable de la normalización en el ámbito europeo dentro del área de ingeniería eléctrica.

[73] La Organización Internacional para la Estandarización es una organización no gubernamental compuesta por representantes de organismos nacionales de normalización de más de ciento cincuenta países.

[74] Véase 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).

adopta decisiones automatizadas con base en cuarenta y ochos preguntas de riesgo y treinta y tres de controles para su mitigación. Funciona como un cuestionario, con factores y parámetros diversos —tipo de decisión, diseño de los sistemas, colectivos afectados e impacto sobre la persona, base de datos que nutre el sistema— que permiten obtener unas puntuaciones de evaluación[75].

De aplicación para el sector público destaca el modelo *Understanding Artificial Intelligence Ethics and Safety* elaborado por el *Alan Turing Institute*, y que se concreta en una guía para el diseño e implementación ética y responsable de los sistemas de inteligencia artificial. Lo hace con base en la definición del ciclo de vida del proyecto de inteligencia artificial y está compuesto por un marco integral de seguridad —resiliencia y robustez— y responsabilidad[76].

En el ámbito del sector privado encontramos el modelo de autoevaluación ética para actores del ecosistema emprendedor del Banco Interamericano de Desarrollo, dirigido a emprendedores y desarrolladores de sistemas de inteligencia artificial del sector privado que, mediante un cuestionario, trata de identificar las áreas de atención principales para prevenir errores, sesgos, discriminación y exclusión como consecuencia del despliegue tecnológico[77]. Parte de un enfoque multidisciplinar que contempla seis dimensiones principales: conceptualización y diseño; gobernanza y seguridad; participación humana; ciclo de vida de los datos y algoritmos; actores relevantes; comunicaciones y participación de terceros.

Finalmente, interesa traer a colación el modelo del *Ada Lovelace Institute*[78], que incorpora dos términos y cuatro enfoques fruto de su estudio de las llamadas cajas negras[79]. Los dos primeros son la auditoría de algoritmos y la evaluación del impacto

---

[75] Véase 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).

[76] Véase Leslie, D. (2019). «Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector». Disponible en Internet: https://zenodo.org/record/3240529/files/understanding_artificial_intelligence_ethics_and_safety.pdf?download=1 (última consulta el 9 de julio de 2022).

[77] Rosales Torres, C. S. *et. al.* (2021). «Autoevaluación ética de IA para actores del ecosistema emprendedor: Guía de aplicación». Disponible en Internet: https://publications.iadb.org/publications/spanish/document/Autoevaluacion-etica-de-IA-para-actores-del-ecosistema-emprendedor-Guia-de-aplicacion.pdf (última consulta el 9 de julio de 2022).

[78] Ada Lovelace Institute (2020). «Examining the Black Box: Tools for assessing algorithmic systems». Disponible en Internet: https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf (última consulta el 9 de julio de 2022).

[79] Cuando hablamos de cajas negras o *black boxes* nos referimos a un tipo concreto de algoritmos basados en inteligencia artificial, entre los que están las potentes y seguras redes neuronales. Se conoce el *input* —lo que se aporta—, es decir, como se configura y prepara la caja, pero se desconocen las razones por las que esta ofrece un *output* concreto —lo que produce, los resultados—. El algoritmo hace predicciones potentes, que mejoran sobremanera los resultados de los sistemas expertos que emulan con reglas lógicas el pensamiento racional o el cálculo humano, pero no sabemos explicar la razón de ello ni cómo lo hacen. Resulta paradigmático en este ámbito las inteligencias artificiales basadas en redes

algorítmico. Para cada uno de estos, se identifican dos enfoques clave. Dentro de la auditoría de algoritmos se encuentra la auditoría de sesgo —con un enfoque específico, no integral, centrado en evaluar los sistemas algorítmicos en busca de sesgos y la inspección regulatoria —con un enfoque amplio, centrado en el cumplimiento de un sistema algorítmico con la regulación o las normas, que requiere una serie de herramientas y métodos diferentes; típicamente realizado por reguladores o auditores profesionales—. Por su parte, la evaluación del impacto algorítmico incorpora la evaluación del riesgo algorítmico —se analizan los posibles impactos sociales de un sistema algorítmico antes de que el sistema se implemente, con monitoreo continuo— y la estimación del impacto algorítmico —evaluación de los posibles impactos sociales de un sistema algorítmico en los usuarios o la población a la que afecta una vez que ya está en uso—.

A la espera de la aprobación del Reglamento europeo sobre inteligencia artificial, en la práctica, los principales actores ya han asumido la vigencia y aplicabilidad del principio de responsabilidad proactiva, que les exige actuar *ex ante* con autoevaluaciones de impacto en derechos fundamentales, con un enfoque amplio y multidisciplinar.

El uso de mecanismos como *toolkits* y *checklists* es en realidad una forma efectiva que permite evaluar con homogeneidad y ciertas garantías distintas fases del diseño y desarrollo tecnológico, así como los procesos que las componen. Con ello se simplifican principios y requisitos complejos[80], se analizan cuestiones que son fundamentales para dar cumplimiento a los principios —como las relativas al ciclo de vida del sistema y de los datos, la posible exposición a sesgos y la definición de los responsables y de los terceros que participan en los procesos— y se validan aspectos cruciales para la formulación, construcción, entreno, implementación y monitorización de cualquier sistema de inteligencia artificial. No podemos en suma valorar de forma negativa[81] ese loable intento de simplificar la complejidad técnica con el noble propósito de conseguir mantener el impacto de los sistemas de inteligencia artificial dentro de un umbral aceptable de riesgo para los derechos fundamentales.

---

neurales aplicadas al ajedrez —*AlphaZero, Stockfish o Leela Chess Zero*—, que siempre ganan no sólo a los humanos sino también a los programas basados en sistemas expertos —humanos o racionales—.

[80] Véase en la misma dirección 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.

[81] Otros autores en cambio critican la deriva por la que se exige un deber de actuación y ponderación en el ámbito de los derechos fundamentales a actores no estatales y empresas privadas de ámbito mundial, por ejemplo, cuando los Estados delegan en las grandes plataformas digitales el control de la publicación de contenidos o las reclamaciones a los motores de búsqueda en el ámbito del llamado derecho al olvido digital. Al respecto, véanse Di Gregorio, G. (2022). *Digital Constitutionalism in Europe Reframing Rights and Powers in the Algorithmic Society*, Cambridge, Cambridge University Press; sobre este mismo problema con el añadido de autoridades administrativas no judiciales ponderando derechos fundamentales en conflicto, Simón Castellano, P. y Dorado Ferrer, X. (2022), cit., 11.

IV. TAXONOMÍA 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, 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[82]. 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[83].

No es posible ni plausible tratar de ofrecer un listado de garantías que 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[84]. Otro buen ejemplo es el caso de la falaz medida de acceso al código íntegro, que podría clasificarse como una acción para facilitar la transparencia y que, sin embargo, no siempre resulta idónea, tal y como sucede 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; por ello, a continuación formularemos una propuesta de clasificación basada en las siguientes categorías: transparencia, explicabilidad, seguridad, participación humana —en el diseño, entreno, implementación y monitorización— y 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

---

[82] Sobre este particular véanse Simón Castellano, P. (2021), cit., 161-202; Presno Linera, M. A. (2022a). «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).

[83] 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 Hueso, L. (2022a), cit., 93.

[84] Véase, sobre la relevancia relativa de las explicaciones en el ámbito de la salud y atención sanitaria, el trabajo de 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.

modo de ver, cada una de las categorías indicadas tiene 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. Publicidad activa y transparencia

Cuando hablamos de transparencia algorítmica nos referimos a la publicidad activa y al derecho de acceso a la información relativa a los algoritmos 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 disponibilidad de la información relativa a los mecanismos de participación humana que estos incorporan[85]. 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[86], 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»[87]. 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.

---

[85] Un estudio completo sobre algunas de las propiedades de la transparencia algorítmica puede encontrarse en el trabajo de Cotino Hueso, publicado con posterioridad a la entrega del presente original, pero cuya referencia se incorpora en fase de revisión por tratarse de una contribución significativa al respeto. Véase Cotino Hueso, L. (2022c). «Transparencia y explicabilidad de la inteligencia artificial y «compañía" (comunicación, interpretabilidad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad…). Para qué, para quién y cuánta», en Cotino Hueso, L. y Castellano Claramunt, J. (Eds.), *Transparencia y explicabilidad de la inteligencia artificial*, Valencia, Tirant lo Blanch, 25-70.

[86] 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, Z. C. (2018). «The Mythos of Model Interpretability». *Queue,* 16, 3.

[87] 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 supuesta proximidad del mismo con otros términos[88], como la explicabilidad, 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 los resultados concretos de los sistemas de inteligencia artificial en términos comprensibles para los seres humanos[89].

Repárese ya en este momento que no es lo mismo entender o comprender porqué un sistema de inteligencia artificial adopta o propone decisiones en un caso concreto, mediante las explicaciones que ofrece el propio algoritmo 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. 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*[90], de tal modo que el llamado «derecho a la explicación de las inferencias razonables»[91] 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[92]. En cambio, la explicabilidad o explicación de una decisión algorítmica se refiere a razones o justificaciones para un resultado concreto o particular, y no así del proceso de decisión en general. Cuando convergen ambas, el resultado es un mayor nivel de eficacia para impugnar o recurrir frente al sistema[93].

---

[88] 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 de Zárate Alcarazo, L. (2022). «Explicabilidad (de la inteligencia artificial)». *Eunomía. Revista en Cultura de la Legalidad*, 22, 328-344, más concretamente, 334-335.

[89] Seguimos la definición propuesta por Doshi-Velez, F. y Kim, B. (2017). «Towards a rigorous science of interpretable machine learning». *arXiv preprint*: 1702.08608.

[90] Sobre los factores para tener en cuenta *ex ante* —explicación local— véase Roig Batalla, A. (2020). *Las garantías frente a las decisiones automatizadas*, Barcelona, J. M. Bosch, p. 203.

[91] Véase 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, y más concretamente, pp. 572 y ss.

[92] Véase Roig Batalla, A. (2020), cit., 75.

[93] 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. Véase 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.

Las subcategorías de la transparencia[94] 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[95]. 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[96], 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 unas 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, o los basados en los tradicionales árboles de decisión, sí cumplen con tal propiedad[97].

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.

---

[94] 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 Arrieta, A. et al. (2020). «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI». *Information Fusion*, 58, 82-115, y más concretamente la tabla 2 de la pág. 90.

[95] 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.

[96] Véase Barredo Arrieta, A. et al. (2020), cit., 87.

[97] Ibídem, 90.

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 auditoria 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[98]. 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 *scraping* analizando únicamente el entrenamiento con pruebas de estrés del sistema[99]. 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 entrenos 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

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

[99] Para más detalles sobre las distintas metodologías de auditorías algorítmicas véase Bernhard Walt, R. V. (2018). «Increasing Transparency in Algorithmic Decision-Making with Explainable AI». *Datenschutz und Datensicherheit*, 10, 613-617, más concretamente pp. 614 y ss.

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[100].

IV.2. La explicabilidad y la falsa sensación de seguridad

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 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[101]. 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[102].

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[103]. 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[104]. 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[105]. La detección se produce en el proceso en el que propio algoritmo

[100] Al respecto, Roig Batalla 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». Roig Batalla, A. (2020), cit., 48.

[101] 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).

[102] Véase Ortiz de Zárate Alcarazo, L. (2022), cit., 334.

[103] Véase Montavon, G. et al. (2018). «Methods for interpreting and understanding deep neural networks». *Digital Signal Processing,* 73, 1-15.

[104] Véase 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

[105] Véase Barredo Arrieta, A. et al. (2020), cit., 83.

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[106]. 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 este último, el algoritmo, 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, 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 Guerrero, las evaluaciones de impacto jueguen un papel relevante para defender un sistema de «legibilidad por diseño»[107]. En función de la tecnología aplicada y del uso concreto, la explicabilidad no es una alternativa real, por inviable[108], y en esos casos habrá que apoyarse en otras garantías en función de los resultados de las autoevaluaciones de impacto[109].

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[110] 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 estas explican sus decisiones en términos humanos. 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 explicables pero también falibles fórmulas decisorias humanas.

IV.3. Seguridad y trazabilidad

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

[106] Véase al respecto Athey, S. y Imbens, G. (2015). «Machine Learning Methods for Estimating Heterogeneous Causal Effects». *arXiv: 1504.01132*

[107] 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 Guerrero, M. (2022), cit., 170.

[108] Véase Edwards, L. y Veale, M. (2017), cit., 65 y ss.

[109] Lo importante «no es tanto el reconocimiento de un sólido derecho a la explicación como avanzar hacia un sistema de algoritmos interpretables». Medina Guerrero, M. (2022), cit., 170.

[110] Véase Russell, S. J. y Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. New Jersey, Prentice Hall.

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 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[111]—, 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[112].

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[113].

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[114]. 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—.

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

[111] Véase el trabajo de Adadi, A. y Berrada, M. (2018). «Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)». *IEEE Access*, vol. 6, 52138-52160.

[112] En la misma dirección véase Barredo Arrieta, A. et al. (2020), cit., 108.

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

[114] Para más detalles véase el interesante trabajo de Kroll, J. A. (2021). «Outlining traceability: A principle for operationalizing accountability in computing systems». *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 758-771.

mejora continua, la efectividad de los controles y la salud del sistema en la práctica, con propuestas de acciones de mejora a futuro[115]. 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. Participación 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[116].

Hasta el momento nos hemos centrado en argumentar qué tecnología —aquellas que estamos dispuestos a aceptar, delimitando ciertas líneas rojas—, para qué —un debate centrado en los usos éticos y plausibles— y cómo —reuniendo ciertas propiedades alternativas como la transparencia, la explicabilidad y la seguridad—. Responder esas cuestiones nos ha llevado hasta aquí y, sin embargo, por sí solas no permitirían evitar situaciones ambivalentes como las vividas en los casos COMPAS y BOSCO, descritos más arriba, en los que la imposibilidad de impugnar una decisión automatizada derivaba de una falta de transparencia y explicabilidad que, a su vez, tenia su origen en un problema mayor.

En el caso COMPAS, se trataba de una empresa privada que era propietaria del código sobre el que existían derechos de propiedad industrial. En el caso BOSCO, desarrollado por el Gobierno, se llegó a conclusión que la revelación relativa de la fórmula algorítmica, además de problemas de propiedad intelectual, podría afectar la seguridad pública y la defensa nacional.

El debate sobre si el desarrollo tecnológico debe ser público o privado, en relación con las decisiones automatizadas que puedan adoptarse en el sector público,

[115] Véanse las normas UNE-ISO cfr. nota 117.

[116] 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 Castellano, P. (2021), cit., 203-207 y 209-213.

es necesario[117], aunque en los casos referenciados anteriormente la respuesta final haya sido en ambos supuestos contra *cive*, esto es, a favor de mantener una opacidad insufrible de la que deriva una clara vulneración del derecho de defensa o al recurso. Sin embargo, parece evidente que en el caso COMPAS, si la propiedad o la titularidad de ciertos derechos sobre la herramienta —poderes de control específicos en relación con los procesos de auditoría y monitorización— hubieran estado en manos de las autoridades, difícilmente se hubieran producido esas consecuencias nefastas para los derechos de los individuos, al menos, en países europeos de tradición civilista.

De nuevo, no debemos perder de vista que el legislador europeo avanza hacía un modelo de corregulación más que de autorregulación —a diferencia de los «mecanismos de autorregulación transparentes» a los que se refiere la Carta de derechos digitales[118]—. 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 LIA—, cuando los algoritmos se empleen por parte del sector público para la toma de decisiones automatizadas, se confie 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 y qué modelos de participación humana son más adecuados para que el legislador europeo o nacional los establezca *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 nivel de participación humana. 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,

[117] Se ha llegado a señalar, incluso, la necesidad de que los algoritmos estén debidamente publicados como normas jurídicas o la exigencia de que existan mecanismos de recurso directo e indirecto frente a estos, equiparando su naturaleza jurídica a la de los reglamentos, al considerar que estos reglan y predeterminan la actuación de los poderes públicos. Véase al respecto Boix Palop, A. (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.

[118] 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).

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.

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[119]. Aquí se trataría de ir un poco más allá, estableciendo por Ley la exigencia de controles reforzados —aunque relativos, puesto que la participación humana por sí sola tampoco es garantía de una inteligencia artificial más segura o potente— 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 protección

[119] 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).

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[120] 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»[121], 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»[122].

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

[120] Véase Roig Batalla, A. (2021), cit., 237-238.

[121] Shahriari, K. y Shahriari, M. (2017). *IEEE standard review — Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 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).

[122] Ibídem.

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 LIA, 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»[123]; en Canadá encontramos el *Advisory Council on Artificial Intelligence*[124] 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 LIA, 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 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[125] 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.

---

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

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

[125] 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.

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[126] 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 sus derechos fundamentales, que puedan derivarse del uso de sistemas de inteligencia artificial»[127].

V. CONCLUSIONES

Como cualquier tecnología, los sistemas de inteligencia artificial a menudo distribuyen beneficios y perjuicios de manera desigual, y también agravan o perpetúan condiciones sociales injustas preexistentes. Sin embargo, la tecnología es neutra[128] y su valoración dependerá del uso que se le atribuya; incluso la premisa anterior —los algoritmos multiplican los sesgos humanos— puede leerse de forma positiva, puesto que la inteligencia artificial podría emplearse, también, con fines terapéuticos, esto es, para descubrir realidades discriminatorias desconocidas y, posteriormente, proceder a corregirlas[129].

A lo largo de este trabajo hemos analizado qué tecnología —delimitando ciertas líneas rojas, con una prohibición general de decisiones automatizadas en el sector público y la definición de técnicas inadmisibles para determinados usos, en sintonía con la LIA—, para qué —exigiendo un debate centrado en los usos éticos y plausibles—, quién —garantías subjetivas y participación humana— y cómo —salvaguardas relativas y alternativas como la transparencia, la explicabilidad y la seguridad, así como las garantías institucionales—.

---

[126] 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.

[127] Ibídem. En el momento de la publicación de este artículo ya se ha publicado que el citado órgano tendrá su sede en la ciudad de A Coruña.

[128] Una cosa bien distinta es que los desarrolladores proyecten sesgos en el código algorítmico.

[129] Como sucedió en el caso de Amazon, que fruto del diseño y aplicación de un algoritmo para la selección de personal descubrió la existencia de un sesgo de género. Véase Simón Castellano, P. (2021), cit., 208.

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, motivo por el cual la solución pasa necesariamente por adscribir la respuesta del Derecho 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.

A lo largo del estudio, además, se ha observado como predomina en la teoría y en la práctica una visión subjetivista del derecho de protección de datos personales, de tutela administrativa, así como de otros derechos, lo que los convierte en ineficaces para los sistemas de inteligencia artificial, en la medida que con independencia de que puedan afectarse bienes jurídicos concretos en cada supuesto específico, estamos hablando de desafío para ciertos intereses supraindividuales, difusos o colectivos de muy difícil tutela. El enfoque propuesto en este artículo, basado en la dignidad humana, el libre desarrollo de la personalidad y la dimensión objetiva de los derechos fundamentales debe permitir superar esa visión individual de ciertos derechos y, también, el peculiar fenómeno de proliferación de vacuas cartas o declaraciones de principios éticos sin carácter normativo[130].

Segunda.—El recurso a la normativa relativa a la transparencia o al derecho de acceso a la información en el sector público también resulta claramente insuficiente, siendo la transparencia una garantía más, integrada por diferentes sub-propiedades interrelacionadas con otras garantías como la explicabilidad o la seguridad. La relevancia de las citadas propiedades depende de cada caso concreto, en función de la naturaleza y el contexto de la tecnología, así como del uso que se le pretende dar. 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, cuya lectura aislada no aporta nada en el marco de las autoevaluaciones de impacto, que permite medir niveles de riesgo para establecer controles de mitigación a futuro con la efectividad deseada. El conjunto de las garantías jurídicas estudiadas pretende, en definitiva, un objetivo mayor: asegurar que ningún ciudadano sea objeto de una decisión automatizada sin poder recurrir, replicar o defenderse frente a tal decisión.

Tercera.—Que el debate esté centrado en los usos y en la tecnología ha permitido al legislador europeo esbozar, en la LIA, determinadas líneas rojas, tanto por lo que se refiere a tecnologías invasivas como en lo relativo a las decisiones automatizadas de las que se derivan efectos jurídicos para las personas. Reivindicar un esfuerzo de detalle mayor al regulador en este ámbito es, en realidad, poco sensato e innecesario.

[130] Sin ir más lejos, la citada Carta de Derechos Digitales.

Lo primero en la medida que resulta muy difícil tratar de delimitar con claridad y detalle fines y tecnologías prohibidas, cuando estamos sumergidos en una vorágine de crecimiento exponencial de la capacidad de procesamiento, con descubrimiento de nuevas técnicas y en la carrera para la computación cuántica, que todo lo cambiará; lo segundo, si tenemos en cuenta las funciones —de inspección, sanción y consultivas— que van a desempeñar ciertas autoridades nacionales y europeas —garantías institucionales—, tales como el Comité Europeo de Inteligencia Artificial y la Agencia Española de Supervisión de la Inteligencia Artificial, que podrán acabar de delimitar la respuesta en el caso concreto.

Cuarta.—Los principios y garantías jurídicas deben reinterpretarse, incluso con fines de obtener cierta eficacia de los derechos, frente a la realidad algorítmica. Defendemos la necesidad de rechazar la tentación de sumarse al creciente reconocimiento de nuevos derechos, que muchas veces se formula mediante instrumentos que no tienen valor normativo o, cuando lo tienen, no van acompañados de garantías y suponen un reconocimiento meramente semántico, y reivindicar la defensa de los derechos fundamentales ya existentes teniendo en cuenta conceptos, categorías y técnicas de protección de derechos e intereses supraindividuales, colectivos, reinterpretados ante el reto tecnológico. Este trabajo responde a un esfuerzo sistematizador, que incorpora una propuesta de taxonomía de las garantías frente a los sistemas de inteligencia artificial, otorgando un papel clave a los modelos de análisis de riesgos y, en concreto, a las llamadas autoevaluaciones de impacto, que proliferan incluso antes de la aprobación de la LIA, en la medida en que a nivel internacional se acepta el principio de responsabilidad proactiva —*accountability*—, con modelos de corregulación basados en la gobernanza algorítmica. En esta misma dirección, el rol de las autoridades de control será clave para identificar grupos de personas potencialmente afectadas por los riesgos e integrar en los procesos de asesoramiento a expertos, técnicos, académicos, auditores externos, etc.

Title:
New Law, new guarantees. a proposal for the reinterpretation of the legal principles in the light of the algorithmic reality.

Summary:
(FALTA)

© UNED. *Revista de Derecho Político*
N.º 117, mayo-agosto, 2023, págs. 153-196

Resumen:

En el presente artículo se estudia el enorme desafío que proyectan los sistemas de inteligencia artificial para los derechos de las personas, planteando los términos del debate con una propuesta que responde a cuestiones clave —qué, para qué, quién y cómo— y que se concreta en la definición de la taxonomía de las garantías frente al empleo de algoritmos para la toma de decisiones que producen efectos jurídicos para los individuos.

Ante la insuficiencia de una respuesta basada en el marco jurídico de la protección de datos, y las limitaciones de un enfoque centrado exclusivamente en la transparencia y el derecho de acceso, el autor defiende la necesidad de encontrar el sustento para medidas y garantías sistémicas en la dignidad humana, el libre desarrollo de la personalidad y la dimensión objetiva de los derechos fundamentales. Una lectura más amplia, que escapa a los derechos individuales y que tiene en cuenta la dimensión social o colectiva de la problemática, en la misma línea que la propuesta europea de reglamento armonizado sobre la materia.

El debate debe centrarse, a juicio del autor, en torno a los usos plausibles, posibles, y las garantías vinculadas a su empleo, aceptando un modelo de corregulación basado en el principio de responsabilidad proactiva y las autoevaluaciones de impacto, que se imponen en la práctica por pragmatismo, con el fin de obtener cierta eficacia y mitigar riesgos en el diseño, entrenamiento, empleo y monitorización de herramientas tecnológicas.

La propuesta formulada en el trabajo incorpora hasta cinco categorías —transparencia, explicabilidad, seguridad, garantías subjetivas y garantías institucionales— que, a su vez, están integradas por distintas propiedades y subcategorías, y cuya eficacia o relevancia depende de cada caso concreto —naturaleza, contexto y alcance de la tecnología y de su empleo—.

Abstract:

This article studies the enormous challenge projected by artificial intelligence systems for the rights of people, proposing the terms of the debate with a proposal that responds to key questions —what, why, who and how— and that is specified in the definition of the taxonomy of guarantees against the use of algorithms for decision-making that produce legal effects for individuals.

Given the insufficiency of a response based on the legal framework of data protection, and the limitations of an approach focused exclusively on transparency and the right of access, the author defends the need to find support for systemic measures and guarantees in the human dignity, the free development of the personality and the objective dimension of fundamental rights. A broader vision, which goes beyond individual rights and considers the social or collective dimension of the problem, along the same lines as the European proposal for a harmonized regulation on the matter.

The debate should focus, in the author's opinion, around plausible, possible uses and the guarantees linked to their use, accepting a co-regula-

tion model based on the principle of proactive responsibility and self-assessments of impact, which are imposed in the practice by pragmatism, to obtain certain efficiency and mitigate risks in the design, training, use and monitoring of technological tools.
The proposal formulated in the work incorporates up to five categories —transparency, explainability, security, subjective and institutional guarantees— which, in turn, are made up of different properties and subcategories, and whose effectiveness or relevance depends on each specific case —nature, context and scope of the technology and its use.

Palabras Clave:

Inteligencia artificial, protección de datos, transparencia, explicabilidad, seguridad, dignidad humana, dimensión objetiva de los derechos fundamentales, garantías institucionales, gobernanza algorítmica.

Key words:

Artificial Intelligence, Data Protection, Transparency, Explainability, Security, Human Dignity, Objective Dimension of Fundamental Rights, Institutional Guarantees, Algorithmic Governance.