

DERECHO DIGITAL Y PROPIEDAD INTELECTUAL

Director: Juan Antonio Moreno Martínez

9

PREVENCIÓN Y REPARACIÓN DE DAÑOS OCASIONADOS POR LA DISCRIMINACIÓN ALGORÍTMICA



Dykinson, S.L.

M^a CRISTINA
BERENGUER ALBALADEJO

**PREVENCIÓN
Y REPARACIÓN DE DAÑOS
OCASIONADOS POR LA
DISCRIMINACIÓN ALGORÍTMICA**

COLECCIÓN
DERECHO DIGITAL Y PROPIEDAD INTELECTUAL

DIRECTOR

JUAN ANTONIO MORENO MARTÍNEZ
Catedrático de Derecho Civil de la Universidad de Alicante

COMITÉ EDITORIAL

ISIDORO BLANCO CORDERO
Catedrático de Derecho Penal (Universidad de Alicante)

FERNANDO CARBAJO GASCÓN
Catedrático de Derecho Mercantil (Universidad de Salamanca)

MANUEL DESANTES REAL
Catedrático de Derecho internacional privado (Universidad de Alicante)

JULIAN LÓPEZ RICHART
Profesor Titular de Derecho Civil (Universidad de Alicante)

JUAN JOSÉ MARÍN LÓPEZ
Catedrático de Derecho Civil (Universidad Castilla-La Mancha)

JAVIER PLAZA PENADÉS
Catedrático de Derecho Civil (Universidad de Valencia)

JULIÁN VALERO TORRIJOS
Catedrático de Derecho Administrativo (Universidad de Murcia)

RAQUEL XALABARDER PLANTADA
Catedrática de Propiedad Intelectual (Universitat Oberta de Catalunya)
Director, European Patent Academy

**PREVENCIÓN
Y REPARACIÓN DE DAÑOS
OCASIONADOS POR LA
DISCRIMINACIÓN ALGORÍTMICA**

M^a CRISTINA BERENGUER ALBALADEJO

No está permitida la reproducción total o parcial de este libro, ni su incorporación a un sistema informático, ni su transmisión en cualquier forma o por cualquier medio, sea este electrónico, mecánico, por fotocopia, por grabación u otros métodos, sin el permiso previo y por escrito del editor. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (art. 270 y siguientes del Código Penal).

Diríjase a Cedro (Centro Español de Derechos Reprográficos) si necesita fotocopiar o escanear algún fragmento de esta obra. Puede contactar con Cedro a través de la web www.conlicencia.com o por teléfono en el 917021970/932720407.

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

La presente monografía se enmarca en el Proyecto I+D+I
de Excelencia PROMETEO 2023 (Referencia CIPROM/2022/40)
de la Consellería de Educación, Universidades y Empleo
de la Generalitat Valenciana:
“La nueva era de los algoritmos y la inteligencia artificial
y su tutela jurídica-privada en el marco de la Unión Europea”.

© Copyright by
M^a Cristina Berenguer Albaladejo
Madrid

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

ISBN: 979-13-7006-322-1
Depósito Legal: M-11647-2025
DOI: <https://doi.org/10.14679/4148>

ISBN electrónico: 979-13-7006-410-5

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

*A mis hijos, Cristina y Vicente,
con el deseo de que encuentren su propósito y no lo dejen escapar.*

*A mi marido, Vicente,
por recordarnos que la aptitud suma, pero la actitud multiplica.*

Índice

Capítulo I. El impacto social de los algoritmos y los sistemas de Inteligencia Artificial en la toma de decisiones.....	17
1. INTRODUCCIÓN Y DELIMITACIÓN DEL OBJETO DE ESTUDIO.....	17
2. APROXIMACIÓN A ALGUNOS «CONCEPTOS CLAVE»: INTELIGENCIA ARTIFICIAL, TOMA AUTOMATIZADA DE DECISIONES, <i>MACHINE LEARNING</i> , ALGORITMOS Y DATOS.....	30
2.1. Inteligencia artificial: concepto, características y categorías	30
2.1.1. <i>Concepto</i>	30
2.1.2. <i>Características de los sistemas de IA</i>	38
2.1.3. <i>Categorías de IA</i>	40
2.2. Toma de decisiones automatizadas e inteligencia artificial: ¿distintas denominaciones de un mismo fenómeno?.....	42
2.2.1. <i>Decisiones plenamente automatizadas y decisiones semiautomatizadas</i>	46
2.3. <i>Machine learning</i>: concepto, tipos, elementos y características	53
2.3.1. <i>Concepto y tipos de machine learning</i>	53
2.3.2. <i>Elementos del machine learning: algoritmos y datos</i>	59
2.3.2.1. Algoritmos	59
2.3.2.2. Datos	62
2.3.3. <i>Características de los sistemas de machine learning</i>	66
 Capítulo II. La discriminación algorítmica a grupos vulnerables	 73
1. VENTAJAS Y RIESGOS DEL USO DE SISTEMAS DE IA PARA LA TOMA DE DECISIONES	73
1.1. La discriminación algorítmica como uno de los principales riesgos de las decisiones de los sistemas de IA.....	78

1.2.	Casos que evidencian la discriminación algorítmica a grupos vulnerables	86
1.2.1.	<i>Particular referencia a tres casos de discriminación algorítmica de gran repercusión que han llegado a los tribunales</i>	93
1.2.1.1.	Compas	94
1.2.1.2.	SyRi	97
1.2.1.3.	Frank.....	98
2.	LA DISCRIMINACIÓN ALGORÍTMICA: CONCEPTO, CARACTERÍSTICAS, TIPOS Y CAUSAS	102
2.1.	El derecho a la igualdad de trato y la prohibición de discriminación: concepto y reconocimiento legal	102
2.2.	Tipos de discriminación	107
2.2.1.	<i>Discriminación directa e indirecta</i>	107
2.2.2.	<i>Discriminación por asociación, por error, múltiple e interseccional</i>	111
2.3.	El fenómeno denominado «discriminación algorítmica»	116
2.3.1.	<i>Concepto de discriminación algorítmica</i>	116
2.3.2.	<i>Discriminación algorítmica y sesgo algorítmico</i>	121
2.3.3.	<i>La discriminación algorítmica como discriminación indirecta</i>	125
3.	CAUSAS POR LAS QUE SE PUEDE PRODUCIR LA DISCRIMINACIÓN ALGORÍTMICA	135
3.1.	Consideraciones previas	135
3.2.	Factores que pueden motivar la discriminación de los sistemas de IA	137
3.2.1.	<i>El factor humano</i>	138
3.2.1.1.	En el diseño del algoritmo o la selección de los datos	139
3.2.1.2.	En la interpretación de los resultados y en el uso del sistema.....	141
3.2.2.	<i>Los conjuntos de datos</i>	144
3.2.3.	<i>Las correlaciones o inferencias derivadas del aprendizaje del sistema</i>	149

Capítulo III. Principales desafíos que plantea la discriminación algorítmica

	155
1. INTRODUCCIÓN.....	155
2. PRIMER OBSTÁCULO: DETECTAR LA DISCRIMINACIÓN	157
2.1. Consideraciones previas sobre la opacidad algorítmica.....	157
2.2. Transparencia algorítmica y explicabilidad.....	157
2.3. Opacidad sobre el uso de sistemas algorítmicos de toma de de- cisiones y sobre el contenido de los algoritmos	164
2.4. Opacidad jurídica y técnica del algoritmo	166
2.4.1. La opacidad jurídica	167
2.4.1.1. Sobre el código fuente.....	167
2.4.1.2. Sobre los datos de entrenamiento	172
2.4.2. La opacidad técnica o consustancial al algoritmo	177
2.5. Medidas para hacer frente a los problemas de opacidad alorít- mica en la toma de decisiones a través de sistemas de IA	181
2.5.1. Estado de la cuestión.....	181
2.5.2. El Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, de protección de datos (RGPD) ...	184
2.5.2.1. Exégesis del art.22 RGPD	187
2.5.3. Otras normativas adoptadas a nivel europeo previas al Regla- mento (UE) 2024/1689 (RIA)	208
2.5.4. El Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial	211
2.5.4.1. Obligaciones de transparencia y comunica- ción de información	211
2.5.4.2. Derecho a una explicación.....	217
2.5.5. Medidas adoptadas a nivel interno.....	221
2.5.5.1. Derecho privado español.....	221
2.5.5.2. Derecho público español.....	225
2.5.6. Pronunciamientos judiciales y de órganos administrativos en materia de transparencia y acceso a la información algorítmica ...	234

2.5.6.1.	Dos casos paradigmáticos (y antagónicos) sobre transparencia algorítmica en el sector público: el caso SyRi y el caso Bosco.....	236
2.5.7.	<i>Otras propuestas para hacer frente a la opacidad algorítmica</i>	246
2.6.	Consideraciones finales sobre la necesidad de transparencia y explicabilidad para detectar la discriminación algorítmica.....	249
3.	SEGUNDO OBSTÁCULO: PROBAR LA DISCRIMINACIÓN Y SU CAUSA	251
3.1.	La carga de la prueba de la discriminación y el estándar de prueba	251
3.1.1.	<i>Consideraciones preliminares</i>	251
3.1.2.	<i>La prueba de la discriminación</i>	254
3.1.3.	<i>Los indicios en el ámbito de la discriminación algorítmica.....</i>	260
3.1.4.	<i>Justificación de la discriminación algorítmica.....</i>	273
3.2.	Breve explicación de la problemática que presenta la fijación del nexo causal en el caso de los daños ocasionados por sistemas de IA: la dificultad de su prueba y los mecanismos para mitigarla.....	277
4.	TERCER OBSTÁCULO: DETERMINAR LOS SUJETOS RESPONSABLES	284
4.1.	Problemas encontrados para la atribución de responsabilidad	284
4.2.	Consideraciones a tener en cuenta sobre la atribución de responsabilidad	289
 Capítulo IV. Medidas para hacer frente a los problemas que plantea la discriminación algorítmica (I): la prevención del daño		297
1.	INTRODUCCIÓN: ABORDANDO LOS RIESGOS DE LA IA DE FORMA INTEGRAL E INTERDISCIPLINAR	297
2.	ANÁLISIS DE LAS PRINCIPALES NORMAS JURÍDICAS DIRIGIDAS A PREVENIR LA DISCRIMINACIÓN ALGORÍTMICA Y LOS DAÑOS QUE DE ELLA PUEDAN DERIVAR	301
2.1.	El Reglamento de Inteligencia Artificial (RIA): regulando desde un enfoque preventivo	301

2.1.1.	<i>El riesgo de discriminación de los sistemas en el RIA</i>	302
2.1.2.	<i>Breve referencia a los modelos de IA de uso general</i>	306
2.1.3.	<i>Los sistemas de alto riesgo en el RIA.....</i>	309
2.1.3.1.	Concepto de sistema de alto riesgo	309
2.1.3.2.	Requisitos de los sistemas de alto riesgo	312
2.1.3.3.	Especial referencia a los requisitos relativos a los conjuntos de datos.....	320
2.1.3.4.	Las obligaciones de los sujetos de la cadena de valor	323
2.2.	El Convenio Marco del Consejo de Europa sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho..	325
2.3.	Las medidas de prevención del daño en el RGPD	327
2.3.1.	<i>Las carencias del RGPD para luchar contra la discriminación algorítmica.....</i>	332
2.4.	Las medidas preventivas recogidas en el Derecho antidiscriminatorio	336
3.	EL PAPEL DE LA ÉTICA Y LA TECNOLOGÍA EN LA PREVENCIÓN DE LA DISCRIMINACIÓN ALGORÍTMICA	339
3.1.	Soluciones desde el punto de vista de la ética algorítmica	341
3.1.1.	<i>Formación y educación de las personas que trabajan con sistemas de IA.....</i>	342
3.1.2.	<i>Composición heterogénea de los equipos que diseñan y auditan los algoritmos.....</i>	345
3.2.	Soluciones que pueden adoptarse desde el ámbito tecnológico	351
3.2.1.	<i>Estrategias para evitar sesgos en los conjuntos de datos</i>	354
3.2.2.	<i>Algoritmos para detectar, mitigar y corregir sesgos de otros algoritmos</i>	356
4.	REFLEXIÓN FINAL SOBRE LAS MEDIDAS PREVENTIVAS PARA EVITAR LA DISCRIMINACIÓN ALGORÍTMICA	357

Capítulo V. Medidas para hacer frente a los problemas que plantea la discriminación algorítmica (II): los daños derivados de la discriminación y su reparación.....	359
1. LAS PRINCIPALES PROPUESTAS DE LAS INSTITUCIONES EUROPEAS PARA SUPERAR LOS RETOS QUE LA IA PLANTEA EN EL ÁMBITO DE LA RESPONSABILIDAD CIVIL	359
1.1. Introducción	359
1.2. La Propuesta de Reglamento anexa a la Resolución del Parlamento Europeo sobre responsabilidad civil, de 20 de octubre de 2020	363
1.2.1. <i>Ámbito de aplicación.....</i>	364
1.2.2. <i>Régimen de responsabilidad civil</i>	365
1.2.2.1. Sujetos responsables	365
1.2.2.2. Criterio de imputación de la responsabilidad	369
1.2.2.3. Causas de exoneración de la responsabilidad	372
1.3. Las dos propuestas de Directivas en materia de responsabilidad civil de la Comisión Europea, de 28 de septiembre de 2022....	375
1.3.1. <i>Su conexión con el RIA</i>	375
1.3.2. <i>Relación entre ambas</i>	377
1.3.3. <i>Análisis de la Propuesta de Directiva relativa a la adaptación de las normas de responsabilidad civil extracontractual a la inteligencia artificial.....</i>	382
1.3.3.1. Líneas generales y principales novedades	382
1.3.3.2. Ámbito de aplicación.....	386
1.3.3.3. Criterio de imputación de la responsabilidad y sujetos responsables	387
1.3.3.4. Medidas para aliviar la carga probatoria.....	391
2. LOS DAÑOS EN EL ESCENARIO DE LA IA Y LOS MECANISMOS PARA SU REPARACIÓN	410
2.1. El daño como elemento indispensable de la responsabilidad civil: ¿novedades respecto a los daños derivados del uso de la IA? ...	410

2.1.1.	<i>Daños indemnizables en la Propuesta de Reglamento del PE de 20 de octubre de 2020, en la Propuesta de Directiva de la CE de 28 de septiembre de 2022 sobre responsabilidad civil en materia de IA y en la Directiva 2024/2853 sobre responsabilidad por productos defectuosos: especial referencia al daño moral</i>	414
2.2.	En particular, los daños derivados de la discriminación	420
2.2.1.	<i>La presunción del art. 27 de la Ley 15/2022 integral de igualdad..</i>	421
2.2.2.	<i>El daño moral presunto y su valoración</i>	426
3.	MECANISMOS DE REPARACIÓN DE LOS DAÑOS DERIVADOS DE LA DISCRIMINACIÓN ALGORÍTMICA	432
3.1.	La indemnización por daños y perjuicios causados por la discriminación: ¿límites indemnizatorios?	436
3.1.1.	<i>La importancia de las acciones colectivas en la lucha contra la discriminación algorítmica</i>	440
3.1.2.	<i>Compatibilidad con la acción de responsabilidad del art. 82 RGPD ...</i>	446
3.2.	Nulidad de pleno derecho	453
3.3.	Acción de cesación de la conducta discriminatoria	460
	Bibliografía	463

Capítulo I

El impacto social de los algoritmos y los sistemas de Inteligencia Artificial en la toma de decisiones

1. INTRODUCCIÓN Y DELIMITACIÓN DEL OBJETO DE ESTUDIO¹

Actualmente son innumerables los sectores y actividades donde se están empleando sistemas para la automatización de decisiones basados en inteligencia artificial (en adelante, IA). Por lo que se refiere al ámbito público, se utilizan en el campo de la salud, educación, justicia, ayudas y bienestar social, entre otros. Y en el sector privado son frecuentes, por ejemplo, en banca, empleo, seguros y consumo².

Tales sistemas suelen emplear algoritmos de aprendizaje automático que funcionan como si fueran humanamente inteligentes para realizar tareas concretas, aprendiendo y mejorando a partir de los datos que se le suministran,

¹ Este trabajo es resultado del Proyecto «La nueva era de los algoritmos y la inteligencia artificial y su tutela jurídico-privada en el marco de la Unión Europea», referencia CIPROM/2022/40 (Programa Prometeo para Grupos de Investigación de Excelencia 2023 de la Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital) y del Proyecto TSI-100927-2023-6, financiado por el Plan de Recuperación, Transformación y Resiliencia Europeo, de la Unión Europea-Next Generation EU a través del Ministerio para la Transformación Digital y de la Función Pública.

² Las Estrategias de Inteligencia Artificial adoptadas tanto a nivel estatal como autonómico, describen su potencial en diferentes ámbitos y áreas de actividad entre los que figuran, el ámbito sanitario, la enseñanza y el aprendizaje o la agricultura y el medioambiente. Vid., MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL, *Estrategia española de Inteligencia Artificial 2024*, o la *Estrategia de Inteligencia Artificial de la Comunidad Valenciana 2018*, de la GVA, p. 5. Ejemplos concretos de estos usos o aplicaciones pueden verse en SORIANO ARNANZ, A./SIMÓ SOLER, E., «Machine learning y Derecho: aprendiendo la (des)igualdad», en S. Barona Vilar (ed.) *Justicia algorítmica y neuroderecho: una mirada multidisciplinar*, Valencia, 2021, pp. 186-187.

sin estar expresamente programados o basarse en reglas previamente codificadas³. Así, gracias al procesamiento y combinación de ingentes cantidades de datos personales o no personales, pueden decidir de forma automatizada y con cierto grado de autonomía, por ejemplo, quién merece ser destinatario de un crédito o de una subvención pública, quién es apto para un determinado puesto de trabajo, quién debe acceder a un centro educativo, o qué probabilidades de reincidencia tiene un preso. Salta a la vista, por tanto, que decisiones de indudable trascendencia en la vida de las personas están siendo adoptadas hoy en día por algoritmos que almacenan, procesan y combinan todo tipo de datos, realizando inferencias en base a ellos, sin que en muchas ocasiones ni siquiera seamos conscientes de ello⁴.

La omnipresencia de estos sistemas algorítmicos ha provocado que se hayan empezado a acuñar términos como el de *algocracia o algoritmocracia*⁵, que se utilicen expresiones como *ubicuidad algorítmica*⁶, o que incluso se hable de la *algoritmización del mundo*⁷ o de que nos encontramos inmersos en la edad del

³ BARRIOS ANDRÉS, M., *Manual de Derecho Digital*, Valencia, 2020, p. 59. Para LAZCOZ MORATINOS, G., «Análisis jurídico de la toma de decisiones algorítmica en la asistencia sanitaria», en *La regulación de los algoritmos*, coord. por Gustavo Manuel Díaz González; Alejandro José Huergo Lora (dir.), 2020, p. 284, a grandes rasgos, el *machine learning* se trata de complejos modelos algorítmicos que frente a la programación clásica, tienen la capacidad de aprender en su tarea programada, rutinaria y automatizada, a partir de ejemplos que se le facilitan y además con la capacidad de manejar un número de variables inabordable por la programación clásica.

⁴ Sobre esta idea apunta RIVAS VALLEJO, P., «Sesgos de automatización y discriminación algorítmica», en *Discriminación algorítmica en el ámbito laboral*, Pilar Rivas Vallejo (dir.), 2022, p. 48, que las consecuencias sociales y políticas de basar decisiones importantes en estos sistemas son inestimables, especialmente cuando la intervención humana es escasa, ya que puede condicionar la vida de muchos individuos, destacando en la p. 49, que actualmente las decisiones críticas no se toman en base a datos, sino en base a datos analizados algorítmicamente.

⁵ Emplean estos términos, entre otros, RIVAS VALLEJO, P., «Sesgos de automatización...», *cit.*, pp. 41-43; o RODRÍGUEZ CARDO, I.A., «Decisiones automatizadas y discriminación algorítmica en la relación laboral: ¿hacia un Derecho del Trabajo de dos velocidades?», *Revista Española de Derecho del Trabajo* num.253/2022, BIB 2022\1779, p.4, quien manifiesta que este término proviene del inglés *algocracy* y que fue acuñado por primera vez en ANEESH, A., *Technologically coded authority: The post-industrial decline in bureaucratic hierarchies*, 2002, disponible en <https://web.stanford.edu/class/sts175/NewFiles/Algocratic%20Governance.pdf>; también SOLAR CAYÓN, J.I., «Inteligencia artificial en la justicia penal: los sistemas algorítmicos de evaluación de riesgos», en *Dimensiones éticas y jurídicas de la inteligencia artificial en el marco del Estado de Derecho*, J.I. Solar Cayón (ed.), Alcalá de Henares, 2020, p. 128.

⁶ LAZCOZ MORATINOS, G., «Análisis jurídico de la toma de decisiones algorítmica ...», *cit.*, p. 285.

⁷ Así, LÓPEZ ONETO, M., *Fundamentos para un Derecho de la inteligencia artificial: ¿queremos seguir siendo humanos?*, Valencia, 2020, pp. 27-37, habla de la «algoritmización digital del mundo» y considera que hasta el momento la IA se busca a través de algoritmos que replican ciertos aspectos de la inteligencia humana a la que en ocasiones superan y que el desenvolvimiento de la vida es una continua ejecución de algoritmos (p. 47).

algoritmo y la cuarta revolución industrial, donde la IA ocupa un lugar protagonista junto con el *Big Data* y el Internet de las Cosas⁸.

Con ello se pretende alertar sobre el incremento de decisiones de gran impacto social, con efectos directos en la vida de las personas, que actualmente se adoptan sin control humano⁹. Y no sólo en cuestiones triviales o cotidianas como podría ser la recomendación de una película en una plataforma de *streaming* o qué ruta seguir para llegar del punto A al punto B en el menor tiempo posible, sino en asuntos de gran relevancia como la concesión o denegación de ayudas sociales o la fijación de la prima a pagar en un contrato de seguro. De hecho, para el Comité Económico y Social Europeo (en adelante, CESE), la pirámide de riesgos sobre la que se asienta la *Propuesta de Reglamento del Parlamento Europeo y del Consejo, de 21 de abril de 2021, por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial)* –en adelante, Propuesta de RIA de 2021¹⁰– viene a confirmar la normalización de la *algoritmocracia*, es decir, la sustitución de decisiones humanas por decisiones guiadas por la IA¹¹ y recomienda que determinadas decisiones sigan correspondiendo a las personas, especialmente en ámbitos donde tengan un elemento moral e implicaciones jurídicas o repercusión social, como en los ámbitos judicial y policial, los servicios sociales, la sanidad, la vivienda, los servicios financieros, las relaciones laborales y la enseñanza¹².

⁸ El concepto de «Cuarta Revolución Industrial» fue acuñado por Klaus Schwab (fundador del Foro Económico Mundial en el contexto de la edición del Foro Económico Mundial 2016) quien sostuvo que si la *tercera revolución industrial* es la revolución digital que ha estado en vigor desde mediados del siglo XX y que se caracteriza por una fusión de tecnologías que está difuminando las líneas entre lo físico, lo digital y lo biológico. La *cuarta revolución* está marcada por avances tecnológicos emergentes en una serie de campos, incluyendo *robótica, inteligencia artificial, cadena de bloques, nanotecnología, computación cuántica, biotecnología, internet de las cosas, impresión 3D, y vehículos autónomos*. Vid., SCHWAB, K., «The fourth industrial revolution: what it means and how to respond», accesible en <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/>. Siguiendo la visión de Schwab, presenta una panorámica de las distintas revoluciones industriales, si bien profundizando en la cuarta, LÓPEZ ONETO, M., *op.cit.*, pp. 27-37. También sobre este fenómeno puede verse, KEMP, R., «Artificial intelligence: introduction», *Practical Law*, 2017, p. 2.

⁹ Sobre ellas se pronuncia el GT29 en sus *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679*, estableciendo que las mismas «representan la capacidad de tomar decisiones por medios tecnológicos sin la participación del ser humano» y pueden basarse en cualquier tipo de datos.

¹⁰ En el momento en que estas páginas están siendo redactadas, ya contamos con el Reglamento IA aprobado y publicado en el DOUE de 12 de julio de 2024: *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial*.

¹¹ Así lo pone de manifiesto RIVAS VALLEJO, P., «Sesgos de automatización ...», *cit.*, p. 51.

¹² Vid., *Dictamen del Comité Económico y Social Europeo sobre la «Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión [COM(2021) 206 final - 2021/106 (COD)]*, ap. 3.7.

Esta transformación digital ha conllevado que se considere a los algoritmos como entes que clasifican y predicen nuestro futuro, como si de oráculos se tratase, que son capaces de adivinar lo que ni siquiera nosotros sabemos, queremos o necesitamos¹³. Ejemplo de lo afirmado son, entre otros muchos, los algoritmos de aprendizaje automático capaces de predecir con gran precisión crisis de salud mental antes de que ocurran, a partir del análisis de las historias clínicas electrónicas de miles de pacientes¹⁴ o de las fotos publicadas en redes sociales¹⁵; aquéllos capaces de diagnosticar paradas cardíacas u otras dolencias analizando la voz de la persona que llama a emergencias¹⁶; o los que predicen embarazos adolescentes a partir del análisis de sus datos personales¹⁷.

Si bien el uso de programas informáticos para el procesamiento automatizado de datos no es un fenómeno nuevo, puesto que desde hace años se vienen utilizando tanto por las Administraciones públicas para cruzar información de diferentes bases de datos, como por las empresas privadas para la adopción de múltiples decisiones en el seno empresarial (ej. selección de personal, concesión de créditos, fijación de primas de seguros, etc.), la creciente capacidad computacional de estos sistemas y el aumento de sus funcionalidades, unido a la cada vez mayor disponibilidad de datos, ha permitido su desarrollo y perfeccionamiento, lo que ha conllevado su uso generalizado en prácticamente todos los ámbitos de la sociedad, buscando velocidad, eficiencia y objetividad en la toma de decisiones¹⁸.

¹³ Tal como apunta BARRIO ANDRÉS, M., *op.cit.*, p. 60, los algoritmos de *machine learning* sustituyen causalidad por predicción, lo que significa que están diseñados para «adivinar» y no para «conocer».

¹⁴ Vid., https://www.upf.edu/es/web/e-noticies/inicio/_/asset_publisher/wEpPxsVRD6Vt/content/id/257954891/maximized#.Yr3RvoRBy5c

¹⁵ Vid., <https://www.publico.es/sociedad/instagram-algoritmo-detectar-infelices.html>

¹⁶ Tal como mencionaba la *Comunicación de la Comisión de 25 de abril de 2018 sobre Inteligencia artificial para Europa (COM(2018) 237 final)*, en Dinamarca, la IA ha contribuido a salvar vidas al permitir a los servicios de emergencias diagnosticar paradas cardíacas u otras dolencias analizando la voz de la persona que llama, y en Austria, ha ayudado a los radiólogos a detectar tumores con mayor precisión, al facilitarles la comparación instantánea de las radiografías con una gran cantidad de otros datos médicos.

¹⁷ Se da noticia de este polémico caso acaecido en Argentina, por ejemplo, en <https://www.abc.es/tecnologia/controvertido-inteligencia-artificial-senala-embarazo-adolescente-20230525203617-nt.html?ref=https%3A%2F%2Fwww.google.com%2F>

¹⁸ SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas. Más allá de la protección de datos», *Revista de Derecho Público: teoría y método*, n. 13, 2021, p. 87. Para GARCÍA TERUEL, R.M., «El derecho de daños ante la inteligencia artificial y el *machine learning*: una aproximación desde las recomendaciones del Parlamento Europeo y del Grupo de Expertos de la Comisión Europea», en *Cuestiones clásicas y actuales del Derecho de daños: estudios en homenaje al profesor Dr. Roca Guillaumon*, coord. por Joaquín Ataz López y José Antonio Cobacho Gómez, Vol. 2, 2021, pp. 1010-1011, la IA que usa técnicas de *machine learning* está abierta a un intenso debate doctrinal, no porque esta sea

Precisamente éstas han sido algunas de las ventajas más destacadas de los sistemas inteligentes para la toma de decisiones. Así, como se analizará con más detalle en su momento, su rapidez, eficiencia, reducción de costes de producción y transacción, y la creencia de que ofrecen una mayor imparcialidad y permiten eliminar sesgos o prejuicios humanos, han sido apuntadas como sus bondades más significativas.

No obstante, tales beneficios se ven ensombrecidas por los riesgos que implican para un buen número de derechos fundamentales¹⁹. Así, además de amenazas a la intimidad o a la protección de datos personales, uno de los mayores peligros del uso de estos *softwares* es la adopción de decisiones sesgadas y discriminatorias que, además de afectar a las personas objeto de la decisión, lleven a reproducir y perpetuar estereotipos y desigualdades contra los que se lleva siglos luchando. Dichos riesgos han sido destacados en prácticamente la totalidad de iniciativas, propuestas legislativas y recomendaciones adoptadas en materia de IA a nivel nacional, internacional y fundamentalmente europeo. Baste citar, por el momento, el reciente *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial*—en adelante, RIA—, donde el peligro de que los sistemas de IA puedan perpetuar patrones históricos de discriminación, por ejemplo contra las mujeres, ciertos grupos de edad, las personas con discapacidad o las personas de orígenes raciales o étnicos concretos o con una orientación sexual determinada, es tenido en cuenta para clasificar ciertos sistemas como de alto riesgo²⁰.

una tecnología nueva o emergente sino por su potencial evolución, funcionalidades, uso generalizado y posibles casos de responsabilidad. Para ATAZ LÓPEZ, J., «Daños causados por las cosas: una nueva visión a raíz de la robótica y de la inteligencia artificial», *Càtedra Jean Monnet Working Paper*; 4/2020, p. 18 (disponible en <https://diposit.ub.edu/dspace/handle/2445/169850>), lo que diferencia los sistemas de IA actuales de los anteriores es el impresionante volumen de datos de que ahora se dispone.

¹⁹ La idea de que los sistemas de IA aportan beneficios pero también implican riesgos ha sido puesta de manifiesto de forma constante tanto por las instituciones europeas e internacionales, como por la mayoría de los autores que han tratado el tema en estos últimos años. Consagra esta idea, por ejemplo, el RIA desde sus versiones iniciales hasta la definitiva. Reconoce dicho texto que esta tecnología tiene un enorme potencial en ámbitos tan diversos como la salud, el transporte, la energía, la agricultura, el turismo o la ciberseguridad pero también entraña riesgos que pueden amenazar la seguridad de las personas y los derechos. También el Secretario General de la ONU afirmaba en su *Estrategia en materia de nuevas tecnologías*, Naciones Unidas, 2018, p. 3 y 8/22, que la IA encierra una promesa extraordinaria para el bienestar humano, pero también puede generar más desigualdad y más violencia; ofrece nuevas maneras de cumplir con los Objetivos de Desarrollo Sostenible, pero también puede exacerbar la desigualdad en y entre las naciones, y contribuir al desempleo, asimilar y profundizar los prejuicios. Y aunque estas tecnologías son muy prometedoras, no están exentas de riesgos y algunas inspiran ansiedad e incluso miedo, ya que pueden utilizarse con fines maliciosos o tener consecuencias negativas no deseadas.

²⁰ Vid., por ejemplo, los Considerandos 56 y 57. También en el Considerando 67 se reconoce expresamente que los sistemas de IA «perpetúan y amplifican la discriminación existente, en particular con respecto a las personas vulnerables pertenecientes a determinados grupos, en particular grupos raciales o étnicos».

También la doctrina identifica como especial amenaza derivada del uso de sistemas de IA para la toma de decisiones, los ataques a la privacidad y la discriminación²¹. Contamos cada vez con más evidencias de que tales decisiones no son siempre neutrales ni objetivas, sino que, muy al contrario, el riesgo de llegar a soluciones parciales, sesgadas o discriminatorias es más elevado de lo que *a priori* pudiera parecer. A lo largo de este trabajo daremos buena cuenta de que en la práctica estos casos son más frecuentes de lo que sería deseable. Algunos de los más mediáticos acaecieron hace ya algunos años en el seno de grandes empresas como Apple, Amazon o IBM, cuyos *softwares* demostraron tener prejuicios sexistas o racistas al haber sido entrenados con datos sesgados y tuvieron que ser retirados. Respecto de otros que también han tenido gran repercusión contamos incluso con pronunciamientos judiciales que vienen marcando el rumbo a seguir en la materia. Así, el caso del algoritmo SyRI (*Systeem Risico Indicatie*, o Sistema Indicativo del Riesgo) resuelto por

²¹ Sobre esta idea puede verse la obra de O'NEIL, C., *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, 2017, donde la autora pone de manifiesto que los algoritmos están reproduciendo, cuando no exacerbando, los errores humanos, los prejuicios y las injusticias. También profundiza sobre esta cuestión el estudio de GERARDS, J./ XENIDIS, R., *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law* (Special report European network of legal experts in gender equality and non-discrimination), European Commission, 2020, y el estudio de ORWAT, C., *Risks of Discrimination through the Use of Algorithms. A study compiled with a grant from the Federal Anti-Discrimination Agency*, Berlín, 2020. Exponen ampliamente los peligros que representa la IA para la privacidad, la democracia y la equidad de los sistemas, MANHEIM, K. M., & KAPLAN, L., «Artificial Intelligence: Risks to Privacy and Democracy», *Yale Journal of Law and Technology* 106, vol. 21, 2019, p. 134, quienes consideran que el riesgo de que estos derechos sean vulnerados aumenta en la medida que un sistema de IA desarrolla su automatización y autoaprendizaje. En nuestra doctrina recogen esta idea, por todos, SORIANO ARNAZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 93, GASCÓN MARCÉN, A., «Derechos humanos e inteligencia artificial», en *Setenta años de Constitución Italiana y cuarenta años de Constitución Española*, Pérez Miras (dir.)/Teruel Lozano (dir.)/ Raffiotta (dir.)/ Pia Iadicco (dir.), Vol. 5, 2020 (Retos en el siglo XXI, coord. por S.Romboli), p. 337, GINÈS i FABRELLAS, A., «Sesgos discriminatorios en la automatización de decisiones en el ámbito laboral: evidencias de la práctica», en RIVAS VALLEJO, P. (dir.), *Discriminación algorítmica en el ámbito laboral*, Navarra, 2022, p. 300, SÁNCHEZ CAPARRÓS, M. «Prevenir y controlar la discriminación algorítmica», *Rubinzal Culzoni*, n° 427, 2021. Recuperado de https://www.researchgate.net/publication/358207305_Prevenir_y_controlar_la_discriminacion_algoritmica, p. 6/34, ALAMEDA CASTILLO, M.T., «Reclutamiento tecnológico. Sobre algoritmos y acceso al empleo», *Temas laborales: Revista andaluza de trabajo y bienestar social*, N° 159, 2021, p. 14. Para RIVAS VALLEJO, P., «Sesgos de automatización...», *cit.*, p. 50, los usos de la IA ponen en peligro diversos derechos fundamentales, no solo la intimidad (en el que se centraron los legisladores europeos en el RGPD), sino también, especialmente, la igualdad y no discriminación, sobre todo a partir de la reducción progresiva de la intervención humana. Por su parte, órganos como el GT29 (actual Comité Europeo de Protección de Datos –CEPD–), ya subrayaba en sus «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», el peligro de que tanto la elaboración de perfiles como las decisiones automatizadas perpetuasen los estereotipos existentes y la segregación social y llevasen a discriminación injustificada.

el Tribunal de lo Civil de la Haya en su sentencia de 5 de febrero de 2020; el caso del algoritmo *Frank*, resuelto por el Tribunal Ordinario de Bolonia en sentencia de 31 de diciembre de 2020; o el caso *Loomis v. Wisconsin*, conocido por el Tribunal Supremo de Wisconsin (EEUU), que giró en torno a un sistema inteligente de evaluación de riesgos denominado *COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)*. En todos ellos, se evidenció que el algoritmo (o el uso que se hacía del mismo) discriminaba a ciertos colectivos²².

En suma, aunque a lo largo de estas páginas incidiremos en esta idea, hay que desterrar ya de entrada la falsa creencia de que los algoritmos son instrumentos objetivos que producen siempre resultados objetivos e imparciales. Los algoritmos, al ser diseñados y entrenados por personas, pueden heredar tanto los sesgos y prejuicios de éstas, como los que están intrínsecos en los propios datos que utilizan para funcionar. Con el inconveniente añadido de que, por sus propias características, dichos sistemas algorítmicos no solo acaban reproduciendo estos sesgos, sino que los amplifican estructuralmente, consolidando y expandiendo la discriminación.

Al análisis de los potenciales *riesgos y daños* al derecho a la igualdad de trato y no discriminación derivados del uso de sistemas de IA en la toma de decisiones dedicaremos el presente trabajo.

En primer lugar, se expondrán una serie de conceptos *técnicos* claves para la comprensión de la problemática *jurídica* objeto de análisis. Tras ello, se pondrán de manifiesto las ventajas y riesgos que con carácter general implica el uso de algoritmos en los procesos de toma de decisiones, para después focalizar la atención en la amenaza particular que suponen para el derecho a la igualdad y no discriminación. Tras esto, se explicará qué es la discriminación algorítmica y qué rasgos la caracterizan, cuáles son los factores o causas que pueden dar lugar a este tipo de discriminación, cuáles son los principales desafíos que plantea y qué medidas pueden adoptarse para hacerles frente, tanto *ex ante* como *ex post*, valorando si las respuestas actualmente ofrecidas por la legislación vigente, así como las pergeñadas en las propuestas proyectadas por las instituciones europeas son suficientes o no para superar dichos obstáculos.

Asimismo, debe tenerse muy presente en el análisis del tema objeto de estudio, que en la medida en que los sistemas automatizados de toma de decisiones se basan en el procesamiento de datos –tanto personales como no personales–, nos podremos encontrar con decisiones algorítmicas que, además de ser discriminatorias, se basen en un tratamiento ilícito de datos persona-

²² Para un análisis más detallado de estos casos, nos remitimos al apartado 1.2.1 del Capítulo II de este trabajo.

les²³. Por tanto, podrá ocurrir que una misma operación o decisión adoptada a través de estos sistemas IA infrinja dos (o más) derechos fundamentales²⁴. Por un lado, el derecho a la igualdad y no discriminación cuando la decisión discrimina a la persona por motivos injustos o en base a categorías protegidas y no está justificada objetiva y razonablemente; por otro lado, la protección de datos personales, cuando el uso masivo y tratamiento de datos utilizados por el sistema IA para su funcionamiento afecta a datos de carácter personal contraviniendo la normativa al efecto²⁵.

En esta tesitura, esto es, cuando la decisión automatizada basada en un tratamiento prohibido de datos personales a su vez genera discriminación, será preciso analizar tanto las respuestas ofrecidas desde el marco jurídico antidiscriminatorio como desde la normativa de protección de datos, cuya interconexión o convergencia se hace necesaria para darles solución, debiéndose interpretar las normas de protección de datos a la luz de las disposiciones protectoras del derecho a la igualdad y no discriminación²⁶. Por el contrario, si la decisión automatizada es discriminatoria pero el tratamiento de datos ha sido correcto (bien porque respetaba la normativa y había base legitimadora para

²³ Ponen de relieve esta situación, por ejemplo, ALAMEDA CASTILLO, M.T., *op.cit.*, p. 45 y HERRERA DE LAS HERAS, R., «Protección de datos e inteligencia artificial», en Cruz Blanca/Lledó Benito (coords.), *La robótica y la inteligencia artificial en la nueva era de la revolución industrial 4.0*, 2021, p. 643 y 645. Para la primera, si un algoritmo se diseña con sesgos y además se vulneran las obligaciones en materia de protección de datos, se producirían dos vulneraciones de derechos fundamentales con las correspondientes obligaciones de reparación privadas y también podría entrar en escena el marco sancionador público. También estudia esta duplicidad de infracciones, CASADESÚS RIPOLL, P., *La responsabilidad civil por el uso discriminatorio de los datos personales a través de la inteligencia artificial*, Granada, 2022.

Sobre la posible compatibilidad de la acción indemnizatoria derivada del art. 82 RGPD y otras acciones, como la del art. 27 Ley 15/2022, nos remitimos al apartado 3.1.2 del Capítulo V.

²⁴ En este trabajo nos centraremos en la vulneración del derecho de igualdad y no discriminación. No obstante, no puede desconocerse que, tal como manifiesta la Agencia de los Derechos Fundamentales (FRA), el uso de sistemas de IA compromete a una gran diversidad de derechos fundamentales, como por ejemplo la privacidad y protección de datos personales, o el acceso a la justicia, y debe prestarse cierta atención a todos ellos.

²⁵ Como pone de manifiesto HERRERA DE LAS HERAS, R., «Protección de datos e inteligencia artificial», *cit.*, p. 653, el uso masivo de datos que conlleva la IA y que puede afectar a datos de carácter personal, será difícilmente compatible con el principio de «minimización de datos» del art. 5 RGPD.

²⁶ Para SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación...», *cit.*, p. 21/31, la combinación de ambos marcos normativos ofrece un sistema más completo de protección contra la discriminación algorítmica y por ello, teniendo en cuenta que las medidas dirigidas a regular y controlar todos los riesgos que plantean los algoritmos se están demorando, los operadores técnicos y jurídicos deberían centrarse en la combinación aplicada de los marcos de protección de datos, igualdad y lucha contra la discriminación a fin de lograr un sistema más eficaz que proteja los derechos a la igualdad y la no discriminación contra los peligros generados por la automatización de procesos de toma de decisiones.

el tratamiento, bien porque la misma no era aplicable por no haberse empleado datos personales o estar estos anonimizados)²⁷, o si a pesar de darse un tratamiento ilícito de datos personales, la misma no tiene efectos discriminatorios, sólo habría vulneración de uno de estos derechos fundamentales que obligaría a acudir a la normativa correspondiente para obtener la reparación del daño. Sea como fuere, si la vulneración de dichos derechos implica daños y perjuicios a la persona, habrá que repararlos cuando concurren los presupuestos para ello, que se analizarán en su momento.

Dando un paso más, junto con la normativa vigente para la protección de ambos derechos, examinaremos, como no podía ser de otra forma, la regulación europea sobre IA dirigida a combatir los efectos perjudiciales de este fenómeno. Sólo así podremos ofrecer una panorámica completa de los mecanismos que podrían (y/o deberían) utilizarse para *prevenir* la discriminación algorítmica, así como, en su caso, para *reparar* los daños que la misma provoque a los particulares objeto de la decisión adoptada o recomendada por estos sistemas.

Por lo que se refiere concretamente a la *reparación de los daños* derivados de *decisiones discriminatorias de los sistemas de IA basadas en datos personales*, habría que dilucidar cómo se deberían aplicar o cómo podrían encajar los diversos regímenes de responsabilidad previstos en las regulaciones aplicables: por un lado, la normativa de protección de datos (en concreto, el art. 82 Reglamento General de Protección de Datos –en adelante, RGPD–), por otro lado, la normativa antidiscriminatoria (principalmente, los arts. 25 y siguientes de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación –en adelante, Ley 15/2022 integral de igualdad–), y por último, la regulación que acabe adoptándose –en su caso– en el ámbito europeo sobre responsabilidad civil extracontractual en el ámbito de la IA, así como la ya aprobada normativa europea sobre responsabilidad por productos defectuosos²⁸. Y es que, tal como aseveraba la Comisión en sus últimas propuestas en la

²⁷ Los sistemas de IA solo pueden tratar datos personales cuando se cuente con la base legitimadora para ello, de acuerdo con los arts. 6 RGPD y 6-10 LOPDGGDD, como podría ser el consentimiento del titular de tales datos o la necesidad del tratamiento para la ejecución de un contrato en el que el interesado es parte. Además, para evitar la vulneración del derecho a la protección de datos personales, deben desarrollarse técnicas preventivas que permitan, entre otras cosas, la seudonimización y anonimización de datos, que han sido avaladas por el GT29 en sus «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», p. 36. Como afirma HERRERA DE LAS HERAS, R., «Protección de datos e inteligencia artificial», *cit.*, p. 658, la anonimización de datos es una herramienta ideal para el tratamiento de datos mediante aplicaciones de IA para evitar vulneraciones de dicho derecho.

²⁸ Directiva (UE) 2024/2853 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, sobre responsabilidad por los daños causados por productos defectuosos y por la que se deroga la Directiva 85/374/CEE del Consejo. Publicada en el DOUE núm. 2853, de 18 de noviembre de 2024 y con entrada en vigor el 8 de diciembre de 2024.

materia, estos regímenes de responsabilidad «*se complementan*» con otras políticas de IA de la Comisión basadas en requisitos preventivos normativos y de supervisión destinados a evitar violaciones de derechos fundamentales, como son el RIA, el RGPD, la Ley de Servicios Digitales y la legislación de la UE sobre no discriminación e igualdad de trato²⁹.

Téngase en cuenta que pocos días antes de cerrar este trabajo la Comisión Europea decidió retirar su *Propuesta de Directiva para adaptar las normas de responsabilidad extracontractual en la Inteligencia Artificial (AI Liability Directive)*. Si bien el régimen de responsabilidad que establecía no era tal, como en su momento se explicará, su virtualidad residía en que incorporaba medidas para facilitar la carga de la prueba de la culpa y la causalidad a las víctimas de los sistemas de IA, así como una importante obligación de los desarrolladores y operadores de IA de proporcionar información relevante sobre el funcionamiento de sus sistemas.

No obstante, la retirada de la propuesta no implica que la Unión Europea haya decidido abandonar la regulación del tema de la responsabilidad. Lo que no se sabe, por el momento, es cómo lo abordará finalmente teniendo en cuenta que sus diversos intentos no han llegado a buen puerto. Es probable que la Comisión Europea explore alternativas que pasen por la adaptación de los marcos legales existentes o la elaboración de nuevas propuestas con otros enfoques que equilibren en mayor medida la protección de los derechos de los ciudadanos con la promoción de la innovación tecnológica. Sobre todo, teniendo en cuenta que como se desprende de las pocas noticias publicadas en el momento de redactar estas líneas, una de las razones más poderosas que al parecer han llevado a retirar la propuesta podría estar relacionada con las preocupaciones sobre posibles obstáculos a la innovación tecnológica. Es decir, teniendo en cuenta que Europa ya cuenta con el Reglamento de IA que impone exigentes requisitos para algunos sistemas y numerosas obligaciones para los operadores de esta tecnología, algunos sectores temen que, junto con ello, una regulación demasiado estricta de su responsabilidad, pueda limitar el desarrollo de la IA en Europa y la competitividad de la industria europea en el contexto global.

Mientras tanto, la *Directiva (UE) 2024/2853 sobre responsabilidad por los daños causados por productos defectuosos*, adoptada en octubre de 2024 y ya vigente, aborda en parte las cuestiones de responsabilidad relacionadas con tales productos entre los que se incluyen, ahora, los programas informáticos «independientemente de su modo de suministro o uso, y, por tanto, *con independencia*

²⁹ Vid., la *Propuesta de Directiva para adaptar las normas de responsabilidad extracontractual en la Inteligencia Artificial (AI Liability Directive)*, y la *Propuesta de Directiva para modernizar el régimen de responsabilidad por productos defectuosos*, ambas de 28 de septiembre de 2022.

de si el programa informático está almacenado en un dispositivo, se accede a él a través de una red de comunicaciones o tecnologías en la nube o se suministra a través de un modelo de programa informático como servicio»³⁰, aclarándose en el Considerando 13 que bajo el concepto de «programa informático» podrían incluirse, por ejemplo, «los sistemas operativos, los microprogramas, los programas de ordenador, las aplicaciones o los *sistemas de IA*». Por lo tanto, los sistemas de IA que operan de forma autónoma, sin estar integrados en *hardware* específico, también quedan sujetos a las disposiciones de la Directiva en cuanto a responsabilidad por daños causados por productos defectuosos.

Para finalizar este apartado introductorio, y con la intención de que el objeto de estudio quede perfectamente delimitado, creemos conveniente matizar que nuestra intención es centrarnos en la prevención y reparación de los daños que la discriminación algorítmica puede ocasionar a los particulares derivada de la toma de decisiones individualizadas, es decir, decisiones que tienen efectos directos sobre personas concretas y dependen de su conducta, se basen o no dichas decisiones en perfilados previos de los sujetos.

No obstante, no puede desconocerse que la discriminación algorítmica no sólo puede darse a través de tratamientos personalizados, sino también a través de la perpetuación de estereotipos perjudiciales para las personas pertenecientes a grupos especialmente protegidos que se produce, por ejemplo, a través de los algoritmos empleados por motores de búsqueda que contribuyen a reproducir, perpetuar y exacerbar discriminaciones históricas hacia minorías étnicas y raciales o hacia las mujeres, como resultado de determinadas combinaciones de palabras introducidas en el buscador³¹.

³⁰ Así lo aclara el Considerando 13 añadiendo que, por el contrario, «la información no debe considerarse un producto, por lo que las normas sobre responsabilidad por productos defectuosos no deben aplicarse *al contenido de los archivos digitales, como los archivos multimedia o los libros electrónicos o el mero código fuente de los programas informáticos*». Tampoco debe aplicarse la Directiva, según indica el Considerando 14, a los programas informáticos libres y de código abierto (*open systems*) desarrollados fuera del contexto de una actividad comercial y que se comparten abiertamente. Vid., también el art.4 que contiene la definición de «producto».

³¹ Así lo pone de manifiesto SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 97. Y para ESPÍN ALBA, I., «Sesgos discriminatorios en la toma automatizada de decisiones en la contratación y protección de datos» en *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, coord. por N. Álvarez Lata/J.M. Busto Lago (pr.), 2024, p. 133 «las consecuencias del tratamiento de datos basado en la IA no se limitan a los conocidos problemas individuales de privacidad y de tratamiento de datos personales, sino que alcanzan una dimensión social con la propagación de prejuicios y estereotipos contra grupos de sujetos o su exclusión de bienes o servicios». Sobre estos casos de discriminación, GERARDS, J./XENIDIS, R., *op.cit.*, p. 59; y GINÈS i FABRELLAS, A., *op.cit.*, p. 300, explica con detalle este problema poniéndolo en relación con asistentes inteligentes como Siri, Alexa o Cortana, mediante los que se reproducen estereotipos de género que pasan inadvertidos porque los hemos normalizado. En el Informe de la Unesco *I'd*

Investigaciones recientes advierten sobre el riesgo de codificar de forma inconsciente los sesgos sociales que se encuentran en la web. Por ejemplo, el trabajo publicado por un grupo de investigadores de la Universidad de Boston refleja que los sistemas de *machine learning* tienen sesgos sexistas porque los datos se extraen de Internet donde es frecuente encontrar asociaciones de conceptos que llevan al sistema a establecer correlaciones como *ama de casa-ella* o *genio-él*³². Así, el algoritmo reproduce los sesgos que existen en los registros. Y el estudio realizado por un equipo de la Universidad de Virginia sobre los sistemas de reconocimiento de imágenes, también añade que no es sólo que los conjuntos de datos utilizados para entrenar a estos sistemas de IA contengan un sesgo de género significativo, sino que los modelos entrenados con estos conjuntos de datos amplifican aún más el sesgo existente. En las colecciones de datos de entrenamiento que se usaban para entrenar los sistemas, las imágenes relacionadas con comprar o lavar estaban vinculadas a mujeres, mientras que las de prácticas deportivas se asociaban a los hombres. El programa, que aprendía con estos datos, no sólo reflejaba los sesgos, sino que los incrementaba porque creaba una asociación aún más fuerte en sus etiquetas al detectar la tendencia subyacente y apostar por ella para acertar. Por ejemplo, en un conjunto de datos de entrenamiento dado, se observó que la actividad de «cocinar» tenía un 33% más de probabilidades de asociarse con mujeres que con hombres; a su vez, el modelo entrenado con este *dataset* amplificaba al 68% la disparidad³³.

En tales casos quedan afectados derechos fundamentales y principios democráticos que los poderes públicos tienen la obligación de proteger, pero los daños que se causan son difusos y ni las víctimas ni los causantes son fácilmente identificables, sobre todo en los supuestos en los que el *software* se encuen-

blush if I could (2019), se precisa que no es casualidad que dichos asistentes virtuales tengan nombres y voces femeninas.

³² Vid., TOLGA BOLUKBASI, KAI-WEI CHANG, JAMES ZOU, VENKATESH SALIGRAMA, ADAM KALAI, «Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings», recuperado de <https://arxiv.org/abs/1607.06520>.

³³ Vid., JIEYU ZHAO, TIANLU WANG, MARK YATSKAR, VICENTE ORDONEZ, KAI-WEI CHANG, «Men also like shopping: reducing gender bias amplification using corpus-level constraints», recuperado de <https://scirate.com/arxiv/1707.09457>. Proponen un modelo que reduce la amplificación del sesgo en un porcentaje considerable. Como apunta en otro de sus trabajos, ZOU, J. «Removing gender bias from algorithms», recuperado de <https://theconversation.com/removing-gender-bias-from-algorithms-64721>, un algoritmo entrenado con noticias de Google News que contienen sesgos, amplificará los estereotipos de género, valorando peor a las mujeres en sectores masculinizados. Por ejemplo, un buscador que ha identificado la profesión de «programador» como masculina, priorizará aquellas páginas web que incluyan nombres masculinos por entenderlas más relevantes, situando las páginas web de mujeres programadoras en los últimos puestos. También puede verse CHEONG, M., y otros, «Ethical Implications of AI Bias as a Result of Workforce Gender Imbalance», CIS & The Policy Lab, The University of Melbourne, 2020.

tra abierto y permite la participación de muchas personas no identificadas en su desarrollo³⁴. Es más, aun siendo posible identificar a las víctimas, el daño individual sería de escasa entidad y no motivaría la interposición de una demanda. Todo ello exige soluciones que en ocasiones exceden del ámbito del Derecho privado al que principalmente se circunscribe este trabajo, pero que evidencian sobradamente las dimensiones del fenómeno que se analiza.

Por último, se ha apuntado que la discriminación en el contexto algorítmico no sólo engloba los conceptos de discriminación directa e indirecta elaborados desde el Derecho antidiscriminatorio europeo, sino también otras situaciones en las que el uso del *machine learning* para la toma de decisiones conlleva un trato desigual a una persona respecto a otras, sin estar vinculada a grupos vulnerables, lo que puede ocurrir, por ejemplo, en los casos de discriminación de precios cuando a una persona se la perfila como un *heavy spenders* por un historial de compras que sugiere ese patrón de comportamiento o por el acceso a las compras en línea con dispositivos de una determinada marca³⁵.

En definitiva, no cabe duda de la trascendencia y magnitud de este nuevo fenómeno surgido en la que se ha venido a denominar la «era de los algoritmos». Además, como se irá viendo en estas páginas, sus manifestaciones son numerosas por cuanto se suele dar en sectores muy diversos, públicos y privados. Respecto a este último, es campo abonado para la discriminación el ámbito laboral o el acceso a bienes o servicios, donde adquieren relevancia, por un lado, tanto el denominado «*scoring crediticio*» en la contratación de servicios financieros, como el uso de estos sistemas para la determinación de las primas en la contratación de seguros, y por otro lado, la llamada discriminación de precios o la publicidad dirigida («*targeting*») en base a una elaboración de perfiles previa («*profiling*») ³⁶.

³⁴ SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 98. Para esta autora, la relevancia de estos daños es el resultado de su valor agregado, esto es, del menoscabo a los principios y valores básicos de la democracia.

³⁵ Ejemplo propuesto por ESPÍN ALBA, I., *op.cit.*, p. 147. Esta autora justifica la importancia y atención que la doctrina ha prestado a la vertiente o dimensión pública de la discriminación algorítmica alegando el carácter invasivo de las herramientas como el reconocimiento biométrico o la predicción de comportamientos ilícitos donde el riesgo de discriminación es más elevado.

³⁶ Sobre esta forma de discriminar señala RUBÍ PUIG, A., «Elaboración de perfiles y personalización de ofertas y precios en la contratación con consumidores», *Revista de educación y derecho*, n° 24 (Relaciones contractuales en la era Digital), 2021, p. 15/24, que se trata de un fenómeno habitual que puede definirse como «la comercialización de unos determinados bienes y servicios en condiciones diferentes para diferentes grupos de consumidores, sin que ello responda a las diferencias en las características de los productos o servicios [...]» donde «un consumidor acaba satisfaciendo un precio diferente al que abona otro consumidor por una mercancía idéntica». Para Alfaro Águila-Real, J., «Precios personalizados y discriminación», accesible en <https://almacendederecho.org/precios-personalizados-discriminacion> (último acceso 15 de octubre de 2024), estas diferencias de trato a los consumi-

2. APROXIMACIÓN A ALGUNOS «CONCEPTOS CLAVE»: INTELIGENCIA ARTIFICIAL, TOMA AUTOMATIZADA DE DECISIONES, *MACHINE LEARNING*, ALGORITMOS Y DATOS

Justificada la relevancia del tema, consideramos conveniente explicar, siquiera brevemente, una serie de conceptos técnicos que a nuestro juicio son esenciales para la adecuada comprensión de la problemática jurídica que plantea el uso de sistemas inteligentes en la toma de decisiones. Vaya por delante que no es nuestra intención hacerlo desde una perspectiva excesivamente técnica (puesto que no es éste nuestro ámbito de conocimiento) sino que el objetivo es explicarlos de forma sencilla y sin desvirtuar su naturaleza, para que puedan ser comprendidos por personas que no son expertas en campos como la ingeniería informática o la ciencia de datos.

2.1. INTELIGENCIA ARTIFICIAL: CONCEPTO, CARACTERÍSTICAS Y CATEGORÍAS

2.1.1. Concepto

No hay un concepto universalmente aceptado de IA, ni un consenso político-regulatorio o académico claro de sus características. Por ello, son muchas y variadas las definiciones que pueden encontrarse en la literatura científica sobre dicha tecnología. La dificultad de proporcionar tal definición reside en que existen muchos tipos de IA (tantos como formas de crearlas y entrenarlas a partir de las posibilidades que ofrecen los algoritmos y los datos que los nutren), con capacidades y grados de autonomía distintos, aunque la mayoría suelen basarse en técnicas de aprendizaje automático³⁷. De hecho, hay quien considera que no es posible, y ni siquiera pertinente, dar un concepto unita-

dores en relación con el precio ofrecido afectan a la dignidad humana (art. 10 CE) lo que justifica un cierto control de precios en forma de una prohibición de discriminación, sobre todo en relación con los consumidores más vulnerables. Sobre el problema de la publicidad dirigida, destaca MUÑOZ GUTIÉRREZ, C., «La discriminación en una sociedad automatizada: Contribuciones desde América Latina», *Revista chilena de derecho y tecnología*, 10(1), 2021, p. 283, que investigaciones sobre la compañía Facebook, han demostrado que la publicidad selectiva es su principal modelo de negocios (*digital redlining*), y que se centra en prácticas tecnológicas que consolidan actos discriminatorios contra grupos ya segregados por la sociedad.

³⁷ Como disciplina científica, la IA incluye diferentes planteamientos y técnicas como el *machine learning* (aprendizaje automático), el *machine reasoning* (que incluye la planificación, temporización, representación del conocimiento y razonamiento, búsqueda y optimización) y la *robótica* (que incluye el control, percepción, sensores y actuadores así como la integración de otras técnicas dentro del sistema ciberfísico). Así lo explica NAVAS NAVARRO, S., «Sistemas expertos basados en

rio de IA en la medida en que se trata de una tecnología compleja y cambiante que engloba un conjunto de aplicaciones muy heterogéneas³⁸.

Pese a ello, desde el ámbito jurídico al que se circunscribe este trabajo, un buen número de autores han venido utilizando el concepto dado por el *Grupo de Expertos de Alto Nivel en Inteligencia Artificial*, que vino a perfeccionar la primera definición de IA ofrecida por la Comisión Europea, según la cual, «El término “inteligencia artificial” (IA) se aplica a los sistemas que manifiestan un comportamiento inteligente, pues son capaces de analizar su entorno y pasar a la acción –con cierto grado de autonomía– con el fin de alcanzar objetivos específicos»³⁹. Posteriormente, el citado Grupo de Expertos definió estos sistemas en sus *Directrices éticas para una IA fiable* como los «sistemas de software (y en algunos casos también de hardware) diseñados por seres humanos que, dado un objetivo complejo, actúan en la dimensión física o digital mediante la percepción de su entorno a través de la obtención de datos, la interpretación de los datos estructurados o no estructurados que recopilan, el razonamiento sobre el conocimiento o el procesamiento de la información derivados de esos datos, y decidiendo la acción o acciones óptimas que deben llevar a cabo para lograr el objetivo establecido»⁴⁰.

El motivo de que se opte mayoritariamente por esta definición quizá sea que en ella se recogen algunas características que permiten calificar un sistema de «inteligente». Entre ellas, la capacidad para interpretar correctamente datos externos y aprender de ellos y emplear después los conocimientos adquiridos para realizar tareas y conseguir objetivos concretos⁴¹; y todo esto, además, con

inteligencia artificial y responsabilidad civil. Algunas cuestiones controvertidas», *Diario La Ley*, 13 de Diciembre de 2019, LA LEY 14980/2019, p. 1/16.

³⁸ Así lo explica ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial y responsabilidad civil*, 2022, p. 37 y ss., citando entre los autores que rechazan que deba darse un concepto unívoco de IA a Bertolini y Palmerini (vid., nota 41). Explica BERTOLINI, A., «Artificial Intelligence and Civil Liability», Study Requested by the European Parliament's Committee on Legal Affairs, 2020, pp. 15-33, que la IA se ha convertido en un término que está de moda (*buzzword*), que es muy utilizado por legisladores, empresas y medios de comunicación y que esto ha llevado a la percepción errónea de que IA es una noción homogénea y fácilmente definible. Sin embargo, en su opinión, es difícil proporcionar una definición completa y unitaria de la IA.

³⁹ COMISIÓN EUROPEA, «Comunicación de la Comisión al Parlamento Europeo, al Consejo Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones “Inteligencia artificial para Europa”», COM/2018/237 final, p. 1. Sigue diciendo esta definición que «Los sistemas basados en la IA pueden consistir simplemente en un programa informático (p. ej. asistentes de voz, programas de análisis de imágenes, motores de búsqueda, sistemas de reconocimiento facial y de voz), pero la IA también puede estar incorporada en dispositivos de hardware (p. ej. robots avanzados, automóviles autónomos, drones o aplicaciones del internet de las cosas)».

⁴⁰ *Directrices éticas para una IA fiable del Grupo Independiente de Expertos de Alto Nivel sobre IA de 2019*, p. 48.

⁴¹ ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...cit.*, p. 40.

cierto grado de autonomía⁴². Asimismo, se trata de una definición lo suficientemente flexible como para adaptarse al progreso continuo que implica esta técnica, y lo suficientemente precisa para proporcionar la seguridad jurídica que debe tener cualquier definición legal, exigencias puestas de relieve por la Comisión Europea en su *Libro Blanco sobre la inteligencia artificial. Un enfoque orientado a la excelencia y confianza* (en adelante, Libro Blanco sobre IA)⁴³.

⁴² En la *autonomía* de estos sistemas también inciden las *Resoluciones del Parlamento Europeo de 20 de octubre de 2020, sobre un marco de los aspectos éticos de la IA, robótica y tecnologías conexas* (2020/2012(INL)), así como la relativa a un *régimen de responsabilidad civil en materia de inteligencia artificial* (2020/2014(INL)). La primera dispone en su art. 4.a) que se trata de «un sistema basado en programas informáticos o incorporado en dispositivos físicos que manifiesta un comportamiento inteligente al ser capaz, entre otras cosas, de recopilar y tratar datos, analizar e interpretar su entorno y pasar a la acción, con cierto grado de autonomía, con el fin de alcanzar objetivos específicos»; y en el apartado b) define la «autonomía» como un sistema de inteligencia artificial que funciona interpretando determinados datos de entrada y utilizando un conjunto de instrucciones predeterminadas, sin limitarse a ellas, a pesar de que el comportamiento del sistema esté orientado a cumplir el objetivo que se le haya asignado y esté limitado por este y por otras decisiones de diseño pertinentes tomadas por su desarrollador». En sentido casi idéntico, los arts. 3 a) y b) de la segunda Resolución citada.

También destaca el rasgo de la autonomía y pone de manifiesto que ha aparecido en la definición de la IA de todos los documentos de trabajo, resoluciones y propuestas del Parlamento Europeo y de la Comisión, MUÑOZ VELA, J.M., *Derecho de la Inteligencia Artificial. Un enfoque global de responsabilidad desde la ética, la seguridad y las nuevas propuestas reguladoras europeas*, Tesis Doctoral, Valencia, 2021, p. 516. Para MARTÍN CASALS, que un sistema sea autónomo significa que, al menos hasta un cierto grado, ante posibles alternativas de actuación toma sus propias decisiones, sin que la elección de una opción específica está determinada por los datos proporcionados por su diseñador. «El sistema de IA solo será “autónomo” cuando, al tomar una “decisión” entre dos o más acciones posibles, se base en sus propios datos». Apunta también este autor que para que un sistema de IA pueda considerarse autónomo, además de gozar de dicha capacidad de decisión, debe tener capacidad de aprender sobre la base de la experiencia y la observación de los datos que recoge (capacidad de aprendizaje), habilidad para adaptarse a los recursos existentes (adaptabilidad), capacidad de poder anticiparse a los acontecimientos de su entorno (previsibilidad) y habilidad de comunicarse con personas y con otras máquinas y cooperar con ellas (capacidad de cooperación). Vid., MARTÍN-CASALS, M., «Desarrollo tecnológico y responsabilidad extracontractual. A propósito de los sistemas de inteligencia artificial (IA)», en J.M. Pérez Collados (Coord.), *La cultura jurídica en la era digital*, Cuadernos Digitales. Derecho y Nuevas Tecnologías, 2022, p. 119. Y según explica con claridad HUERGO LORA, la idea de autonomía es fundamental para distinguir los sistemas de IA de las aplicaciones informáticas «regladas», ya que en los primeros el sistema no se limita a realizar operaciones a partir las reglas que previamente se le han dado y de los datos de entrada (como una calculadora, o como la aplicación «Bosco»), sino que “*pone algo de su parte*”, incluso con “*capacidad de adaptación tras el despliegue*”. Por tanto, estos sistemas sí influyen sobre el contenido de la decisión». Vid., HUERGO LORA, A., «Por qué aciertan las sentencias sobre el ‘algoritmo’ del bono social eléctrico», *Almacén del Derecho*, 10 de mayo de 2024.

⁴³ Así lo considera, por todos, SÁNCHEZ CAPARRÓS, M., *op.cit.*, pp. 2-4. Ambas exigencias también se recogen en el *Libro Blanco sobre la inteligencia artificial. Un enfoque orientado a la excelencia y confianza* (2020), p. 20, donde se afirma que la definición que se otorgue de la IA tendrá que ser suficientemente flexible para adaptarse al progreso técnico al tiempo que mantiene un nivel de precisión adecuado para ofrecer la seguridad jurídica necesaria. Disponible en <https://op.europa.eu/es/publication-detail/-/publication/ac957f13-53c6-11ea-aece-01aa75ed71a>

Otras definiciones, como la proporcionada por la R.A.E., apuntan que la IA es la capacidad de una máquina de *actuar como lo hace una mente humana*, considerando sus aspectos de creatividad y capacidad de realizar análisis e inferencias a partir de información compleja e incluso incompleta⁴⁴. Este rasgo ya se destacó la primera vez que se definió el término «Inteligencia Artificial», en 1956, por John McCarthy, quien dijo que se trataba de «la ciencia e ingeniería de hacer máquinas que se comporten de una forma que llamaríamos inteligente si el humano tuviese ese comportamiento»⁴⁵. No obstante, existe controversia doctrinal acerca de si la «inteligencia» puede predicarse de una máquina o, por el contrario, sería exclusivamente atribuible a los seres humanos⁴⁶.

Especial mención merece la definición finalmente acogida en el RIA, por cuanto pretende armonizar la materia a nivel supranacional. En su primera versión se definía el sistema de IA como «*el software que se desarrolla empleando una o varias de las técnicas y estrategias que figuran en el anexo I y que puede, para un conjunto determinado de objetivos definidos por seres humanos, generar información de salida como contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa*»⁴⁷.

Se explicaba en los Considerandos de la Propuesta que dicha definición pretendía «ser lo más tecnológicamente neutra posible y resistir al paso del

⁴⁴ La RAE define esta tecnología como la «Disciplina científica que se ocupa de crear programas informáticos que *ejecutan operaciones comparables a las que realiza la mente humana*, como el aprendizaje o el razonamiento lógico» (la cursiva es nuestra). En esta línea, SALAZAR GARCÍA, I., define la IA de forma simplificada como un *software*, es decir, un programa informático, dotado de una serie de peculiaridades que lo asemejan a la inteligencia humana, por ejemplo, tiene la capacidad de tomar decisiones. En <https://www.noentiendonada.es/idoia-salazar-la-ia-puede-hacer-mucho-mal/>.

⁴⁵ El término fue acuñado por dicho autor durante una conferencia en Dartmouth College, donde se organizó el *Dartmouth Summer Research Project on Artificial Intelligence*. McCarthy, junto con otros investigadores como Marvin Minsky, Allen Newell y Herbert A. Simon, definió la IA como el campo dedicado a crear máquinas que puedan simular la inteligencia humana. Vid., MCCARTHY, J./MINSKY, M./ROCHESTER, N./SHANNON, C., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. 31.08.1955, Dartmouth College, 1956. Recuperado de: <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html> (consultado el 15 de octubre de 2024). Sin embargo, los conceptos e ideas que estaban detrás de la IA son anteriores puesto que, en la década de 1940, los matemáticos Norbert Wiener y John von Neumann, ya estaban trabajando en la teoría de los sistemas y la computación que sentaron las bases de la IA.

⁴⁶ Vid., sobre las distintas opiniones al respecto, ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...*cit., pp. 43-46.

⁴⁷ El Anexo I establecía que estas técnicas y estrategias eran las siguientes: «a) Estrategias de aprendizaje automático, incluidos el aprendizaje supervisado, el no supervisado y el realizado por refuerzo, que emplean una amplia variedad de métodos, entre ellos el aprendizaje profundo. Estrategias basadas en la lógica y el conocimiento, especialmente la representación del conocimiento, la programación (lógica) inductiva, las bases de conocimiento, los motores de inferencia y deducción, los sistemas expertos y de razonamiento (simbólico). Estrategias estadísticas, estimación bayesiana, métodos de búsqueda y optimización».

tiempo lo mejor posible, habida cuenta de la rápida evolución tecnológica y del mercado en relación con la IA [...]». Sin embargo, pese a los esfuerzos desplegados por la Comisión por dar una definición precisa y perdurable, la misma fue muy criticada por varios motivos⁴⁸: primero, porque a juicio de la doctrina científica, ni todas las técnicas y estrategias recogidas en ese «Anexo» eran IA, ni muchas de las técnicas importantes de IA aparecían en dicho listado⁴⁹; segundo, porque remitir a «una serie de técnicas y estrategias» respecto a una tecnología cuyo desarrollo era exponencial, implicaba su rápida desactualización y el surgimiento de otras nuevas con mayor potencial dañino y no previstas en el Anexo citado⁵⁰.

También se criticó, por un lado, que la definición girara en torno al *software* porque dicho concepto era demasiado amplio y permitía incluir simples programas de ordenador, y por otro lado, que no se aludiera expresamente al rasgo de la *autonomía* que caracteriza a estos sistemas⁵¹. No obstante, como apuntó la doctrina, dicha omisión no significaba que se prescindiera de

⁴⁸ Criticando esta definición, entre otros, TAPIA HERMIDA, para quien la misma resulta «manifiestamente mejorable tanto en términos lógicos como lingüísticos. Vid., <http://ajtapia.com/2022/12/inteligencia-artificial-responsable-iar-la-propuesta-de-directiva-sobre-responsabilidad-en-materia-de-ia-de-28-de-septiembre-de-2022-1/>. Para MUÑOZ VELA, J.M., *op.cit.*, p. 515, la definición dista mucho de lo que debe ser la deseable claridad, concreción y sencillez a la hora de definir un concepto jurídico esencial, técnico y tan complejo. Los motivos se explican en las páginas siguientes.

⁴⁹ Así lo considera el CESE en su *Dictamen del Comité Económico y Social Europeo sobre la «Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión»* [COM(2021) 206 final - 2021/106 (COD)], en el que recomienda la supresión de dicho Anexo por no aportar ningún valor añadido a la propuesta y la modificación del art.3 de forma que la definición de sistema de IA quedara de la siguiente manera: «el *software* que puede, de forma automática y para un conjunto determinado de objetivos definidos por seres humanos, generar información de salida como contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa».

⁵⁰ Así opina, entre otros, ÁLVAREZ CUESTA, H. «Inteligencia artificial: derecho de la UE y derecho comparado. La propuesta de una ley sobre IA», en P. Rivas Vallejo (dir.), *Discriminación algorítmica en el ámbito laboral*, Navarra 2022, p. 382.

⁵¹ EBERS, M.; HOCH, V.R.S.; ROSENKRANZ, F.; RUSCHEMEIER, H.; STEINRÖTTER, B. «The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)», *J—Multidisciplinary Scientific Journal*, 2021, 4, pp. 589–603. <https://doi.org/10.3390/j4040043>. En p. señalan los autores que un enfoque tan amplio puede conducir a una inseguridad jurídica para los desarrolladores, operadores y usuarios de sistemas de IA y que generalmente se asocia el término IA, principalmente con el aprendizaje automático, y no con procesos de automatización simples en los que las reglas preprogramadas se ejecutan de acuerdo con el razonamiento basado en la lógica. Para MUÑOZ VELA, J.M., *op.cit.*, p. 516, la definición es «compleja a nivel técnico y excesivamente genérica y amplia respecto de los sistemas contemplados y no exhaustiva donde el grado de autonomía del sistema desaparece de la definición y se muestra como irrelevante para su calificación como tal, por lo que casi cualquier *software* común con cualquiera de esas capacidades y características sería considerado inteligencia artificial».

ella puesto que, por un lado, seguía estando presente en el Preámbulo de la Propuesta –en concreto en el Considerando 6– donde se reconocía que los sistemas de IA podían diseñarse para operar «con distintos niveles de autonomía», y por otro lado, porque algunas de las técnicas recogidas en el Anexo al que remitía la propia definición, eran precisamente las que emplean los sistemas que actúan con autonomía, como el aprendizaje no supervisado o el profundo⁵².

A la vista de estas objeciones, en la Orientación General del Consejo de 25 de noviembre de 2022 se modifica el art. 3.1 pasando a definirse el sistema de IA como «*un sistema concebido para funcionar con elementos de autonomía que, a partir de datos e información generados por máquinas o por seres humanos, infiere la manera de alcanzar una serie de objetivos, utilizando para ello estrategias de aprendizaje automático o estrategias basadas en la lógica y el conocimiento, y produce información de salida generada por el sistema, como contenidos (sistemas de inteligencia artificial generativa), predicciones, recomendaciones o decisiones, que influyen en los entornos con los que interactúa el sistema de IA*».

Como puede observarse, el concepto ya no gira en torno al *software* sino al «sistema», término que implica algo más y distinto al mero programa informático. Además, se menciona expresamente el rasgo de autonomía con que funcionan dichos sistemas y se concretan las estrategias que pueden utilizar para alcanzar los objetivos fijados, esto es, estrategias de aprendizaje automático o estrategias basadas en la lógica y el conocimiento, eliminándose el Anexo I anteriormente referenciado y los correspondientes poderes de la Comisión para actualizarlo⁵³. Según se explica, con esta modificación que restringe la definición de los sistemas de IA a aquellos desarrollados a través de estrategias de aprendizaje automático y estrategias basadas en la lógica y el conocimiento, se proporcionan criterios suficientemente claros para distinguirlos de otros sistemas de *software* más clásicos⁵⁴.

La misma línea conceptual se ha mantenido en la versión definitiva del RIA, de 13 de junio de 2024, si bien con algunos cambios no demasiado signi-

⁵² Así lo mantiene ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...*cit., p. 41. Para esta autora no tendría sentido que la definición de la IA prescindiera del rasgo de la autonomía puesto que, si así fuera, cualquier tipo de *software* podría considerarse un sistema inteligente, lo que sería ilógico y alejado de la realidad.

⁵³ En su lugar, se añaden los considerandos 6 *bis* y 6 *ter* para aclarar los conceptos de «estrategia de aprendizaje automático» y «estrategias basadas en la lógica y el conocimiento». Para garantizar que el Reglamento de Inteligencia Artificial sea flexible y pueda adaptarse a las transformaciones futuras, se añade, en el art. 4, la posibilidad de adoptar actos de ejecución para especificar y actualizar las técnicas de las estrategias de aprendizaje automático y de las estrategias basadas en la lógica y el conocimiento.

⁵⁴ P. 4.

ficativos. El art. 3.1 define el «sistema de IA», como un «sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que puede influir en entornos físicos o virtuales».

Es el Considerando 12 el que se encarga de desarrollar esta definición. De su lectura se pueden extraer las siguientes ideas:

En primer lugar, que en la misma línea que se venía manteniendo con anterioridad, se pretende proporcionar una definición de estos sistemas que sea, por un lado, lo más clara y *armonizada* posible con la empleada por las organizaciones internacionales que se ocupan de la IA, a fin de garantizar la seguridad jurídica y facilitar su amplia aceptación a nivel internacional; y por otro lado, lo suficientemente *flexible* para dar cabida a los rápidos avances tecnológicos que se produzcan en este ámbito.

En segundo lugar, también se mantiene que la definición debe basarse en las principales características de estos sistemas, como son su capacidad de inferencia, autonomía o adaptabilidad tras su despliegue, rasgos que los diferencian de los *softwares* más clásicos y de la programación tradicional y se excluyen del concepto los «sistemas basados en las normas definidas únicamente por personas físicas para ejecutar automáticamente operaciones».

En tercer lugar, se pone el acento en la capacidad de inferencia de estos sistemas como uno de sus principales rasgos y se explica que, en base a la misma, estos sistemas pueden obtener información de salida y deducir modelos o algoritmos a partir de la información de entrada o datos suministrados. Se destaca, además, que dicha capacidad trasciende el tratamiento básico de datos y permite el aprendizaje, el razonamiento o la modelización. Por tanto, la actividad inteligente es la inferencia o deducción a la que llega el sistema y se realiza de forma automática en base a los datos suministrados cualquiera que sea su origen (obtenidos o no de forma autónoma)⁵⁵.

En cuarto lugar, se pone de relieve que estas inferencias pueden adoptar distintas formas dependiendo de la función que desempeñe el sistema, a saber, predicciones, contenidos, *recomendaciones o decisiones*, pero a diferencia de lo establecido en la versión anterior donde parecía imperativo que dichos resultados influyeran en los entornos con los que interactuase el sistema de IA, parece que ahora se deja abierta *la* posibilidad de que dicha influencia no

⁵⁵ En el texto de 25 de noviembre de 2022, se decía que esos datos a partir de los cuales el sistema funcionaba podían haber sido generados por máquinas o por seres humanos. Vid., Considerando 6.

se produzca, ya que, a tenor del art. 3.1, el resultado arrojado por el sistema «puede influir» en entornos físicos o virtuales, entendiéndose por «entornos» los contextos en los que funcionan los sistemas de IA⁵⁶.

En quinto lugar, se concreta que las técnicas que permiten a los sistemas llegar a dichas inferencias incluyen, por un lado, estrategias de aprendizaje automático en virtud de las cuales los sistemas aprenden de los datos cómo alcanzar determinados objetivos, y por otro lado, estrategias basadas en la lógica y el conocimiento que infieren a partir de conocimientos codificados o de una representación simbólica de la tarea que debe resolverse.

En sexto lugar, se matiza que el término «basado en una máquina» se refiere al hecho de que los sistemas de IA se ejecutan en máquinas;

En séptimo lugar, se explica que los sistemas de IA pueden funcionar con arreglo a objetivos definidos explícitos o implícitos (tareas concretas), y que dichos objetivos pueden ser diferentes de su finalidad prevista en un contexto específico. No obstante, a diferencia de la versión original de la Propuesta, se omite en ésta la mención a que tales objetivos deben ser definidos por seres humanos.

En octavo lugar, se destaca que los sistemas de IA están diseñados para funcionar con distintos niveles de autonomía, lo que significa que pueden actuar con cierto grado de independencia con respecto a la actuación humana y tienen ciertas capacidades para funcionar sin intervención humana. Por tanto, se vuelve a integrar y a mencionar expresamente en la definición el rasgo de la autonomía que había desaparecido, al menos aparentemente, en la versión original del RIA.

En noveno lugar, se explica que la capacidad de adaptación de los sistemas tras su despliegue se refiere a las capacidades de autoaprendizaje que permiten al sistema cambiar mientras está en uso. Es decir, como después explicaremos, al contrario de lo que ocurría en la programación tradicional, en este tipo de sistemas que simulan la inteligencia humana hay aprendizaje porque dependiendo de los datos de entrada que se aporten, el algoritmo genera

⁵⁶ Señalaba VICENTE PALACIO, M.A., «La inteligencia artificial en la ley integral para la igualdad y no discriminación», *Revista General de Derecho del Trabajo y de la Seguridad Social*, N° 64, 2023, p. 135, respecto a la versión original del RIA que «[e]l resultado de esa actividad inteligente puede ser de cualquier tipo (contenido, predicción, recomendación o decisión) siempre que influya en el entorno con el que actúa. Esa influencia no tiene por qué ser necesariamente en su forma más extrema, es decir, aquella que determina la forma de actuar de la persona con quien interactúa (decisión). Entender lo contrario supondría tener que valorar la relación de causalidad entre el contenido producido por el sistema de IA y la posterior actuación humana. Cualquier contenido autónomo producido por estos sistemas a partir del manejo de datos constituye un sistema de inteligencia artificial para la Propuesta de Reglamento».

unas reglas u otras, pudiendo cambiar esas reglas automáticamente mientras está funcionando si, por ejemplo, cambia el contexto social dado (v.gr. guerras, pandemias, etc).

Por último, se manifiesta que los sistemas de IA pueden utilizarse de manera independiente o como componentes de un producto, con independencia de si el sistema forma parte físicamente del producto (integrado) o contribuye a la funcionalidad del producto sin formar parte de él (no integrado).

2.1.2. Características de los sistemas de IA

Por lo que se refiere a las características de los sistemas de IA, junto con las contempladas en la propia definición del RIA, con anterioridad el Grupo de Expertos en responsabilidad civil y nuevas tecnologías (*Expert Group on Liability and New Technologies - New Technologies Formation*, conocido por sus siglas en inglés NTF)⁵⁷, apuntó otras que eran comunes a las tecnologías digitales emergentes en general y que podían complicar la determinación de la responsabilidad civil cuando se ocasionaran daños. En su Informe titulado *Liability for artificial intelligence and other emerging digital technologies* (en adelante, Informe NTF 2019)⁵⁸ señalaba las siguientes: (a) complejidad, (b) opacidad, (c) apertura, (d) autonomía, (e) previsibilidad, (f) manejo y dependencia de datos, y (g) vulnerabilidad⁵⁹.

Respecto a la *complejidad*, señala que el *hardware* actual se compone de múltiples partes que requieren un alto grado de sofisticación técnica para su interacción, y que combinadas con un porcentaje cada vez mayor de compo-

⁵⁷ En su Comunicación al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones «Inteligencia artificial para Europa», 25.4.2018, COM(2018) 237 final, pp. 11 y 17, <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=COM%3A2018%3A237%3AFIN>, la CE anunció la creación del aludido Grupo de Expertos y le encargó la elaboración de un informe en el que se evaluaran los retos que plantean algunas características de la IA respecto de la responsabilidad civil (p. 19).

⁵⁸ El Informe NTF 2019, fue elaborado por 73 expertos independientes de diferentes Estados Miembros, procedentes del mundo académico y empresarial y supone la primera aproximación desde instancias comunitarias al estudio de la responsabilidad por daños derivada de accidentes provocados por tecnologías basadas en IA, como los robots, los asistentes virtuales o los coches autónomos. Para una valoración crítica de este importante informe, puede verse el trabajo de BERTOLINI, A./EPISCOPO, F., «The Expert Group's Report on Liability for Artificial Intelligence and Other Emerging Digital Technologies: a critical assessment», *European Journal of Risk Regulation*, Vol. 12:3, 2021, pp. 644-659, DOI: <https://doi.org/10.1017/err.2021.30>

⁵⁹ Estas características también se recogen por la Comisión Europea en su *Informe al Parlamento Europeo, al Consejo y al Comité Económico y Social Europeo, sobre las repercusiones en materia de seguridad y responsabilidad civil de la inteligencia artificial, el internet de las cosas y la robótica*, de 19 de febrero de 2020 (COM/2020/64 final).

nentes digitales, incluida la IA, conlleva problemas para determinar tanto la causa del daño como su responsable.

Por lo que se refiere a la *opacidad* («efecto caja negra»), se trata de la dificultad para comprender y explicar cómo los sistemas de IA toman sus decisiones⁶⁰.

El rasgo de la *apertura* hace referencia a que los sistemas de IA, por su propia naturaleza, necesitan estar abiertos a actualizaciones y mejoras tras su puesta en circulación, así como interactuar con otros sistemas y fuentes de datos para su correcto funcionamiento. Esta combinación de *hardware* y *software* en continua actualización, tiene un impacto considerable en el régimen de responsabilidad por productos.

En cuanto a la *autonomía* de los sistemas, significa que cada vez tienen mayor capacidad para realizar tareas con poco control o supervisión humana y para tomar decisiones sin estar previamente programados para elegir una opción concreta entre las posibles alternativas de actuación⁶¹. Además, su capacidad de autoaprendizaje conlleva que en función de los datos que se le van suministrando su código fuente pueda cambiar para ajustarse a las nuevas realidades.

Por lo que se refiere a la *previsibilidad* de su actuación, advertía el Informe que cuanto más autónomo fuese el sistema, cuantos más datos externos pudiese procesar y cuanto más sofisticada fuera la IA con la que estuviera equipado, más difícil sería prever su impacto una vez puesto en funcionamiento.

En cuanto al hecho de que estos sistemas no sólo *funcionen con datos*, sino *que dependan de ellos*, señalaba también el Informe que, teniendo en cuenta que la información suministrada no es información preinstalada, sino que se

⁶⁰ Advierte la Comisión Europea en el Informe citado en la nota inmediatamente anterior, que si bien los humanos no tienen por qué comprender todos y cada uno de los pasos del proceso de toma de decisiones, dado que los algoritmos de la IA se van volviendo más avanzados y se van introduciendo en ámbitos críticos, es fundamental que puedan comprender cómo toman las decisiones algorítmicas estos sistemas. Este aspecto reviste especial importancia para el mecanismo de control *a posteriori*, ya que hará posible que las autoridades de garantía del cumplimiento puedan localizar al responsable original de los comportamientos y opciones de los sistemas de IA. Vid., pp. 10-11.

⁶¹ No obstante, como apunta también la Comisión en el Informe citado, (p. 8), si bien los productos basados en la IA pueden actuar de manera autónoma sirviéndose de su percepción del entorno sin seguir una serie de instrucciones predeterminadas, su comportamiento está limitado por el objetivo que se les ha asignado y otras opciones de diseño pertinentes decididas por sus desarrolladores. Téngase en cuenta, además, que actualmente según se desprende del RIA no se permiten los sistemas plenamente autónomos o que tomen decisiones sin ningún tipo de control puesto que se exige que su diseño y desarrollo permitan la supervisión humana. En el art. 14 se establece dicha exigencia y que las medidas de supervisión serán proporcionales a los riesgos, al nivel de autonomía y al contexto de uso del sistema de IA de alto riesgo.

genera por sensores incorporados o comunicada desde el exterior mediante fuentes de datos regulares o proveedores *ad hoc*, es lógico que si los datos suministrados son defectuosos o inexistentes, tanto por errores de comunicación o problemas de la fuente de datos externa, puedan generar resultados incorrectos.

Por último, como consecuencia directa de su diseño abierto, su interconectividad y su dependencia de datos externos, se ponía de relieve su especial *vulnerabilidad* frente a ciberataques que provoquen un mal funcionamiento del sistema o que modifiquen sus características de tal forma que aumenten las probabilidades de causar daños⁶².

2.1.3. Categorías de IA

Para finalizar este apartado, es preciso apuntar que existen varias categorías de IA que se pueden clasificar desde distintos enfoques o en función de diversos criterios como su nivel de autonomía, sus capacidades cognitivas y sus aplicaciones específicas. No obstante, nos limitaremos a explicar una clasificación frecuente hoy en día como es la que distingue entre las inteligencias artificiales *generales*, *fuertes* y *débiles*, a la que se añade la *IA de propósito general*.

La *IA general* se refiere a los sistemas que tendrían la capacidad de resolver cualquier tarea intelectual resoluble por un ser humano. Estos sistemas poseerían una comprensión completa y profunda del mundo, la capacidad de aprender y adaptarse a nuevas situaciones de manera rápida y efectiva, así como la capacidad de razonar sobre un amplio abanico de temas. Por su parte, la *IA fuerte o superinteligencia* iría más allá de las capacidades humanas. Vendría a ser la que vemos en las películas de ciencia ficción donde las máquinas poseen no sólo inteligencia, sino también conciencia. Se trataría de sistemas con capacidad de realizar tareas intelectuales y cognitivas a un nivel que podría superar a los humanos⁶³.

⁶² También en el *Informe de la Comisión al Parlamento Europeo, al Consejo y al Comité Económico y Social Europeo, sobre las repercusiones en materia de seguridad y responsabilidad civil de la inteligencia artificial, el internet de las cosas y la robótica* (COM/2020/64 final), se apunta que la conectividad es una funcionalidad fundamental de un número cada vez mayor de productos y servicios que pone en jaque su seguridad, ya que la conectividad puede, directamente, comprometer la seguridad del producto e, indirectamente, cuando es susceptible de ser pirateado, dar lugar a amenazas para la seguridad y afectar a la seguridad de los usuarios.

⁶³ Así las distingue la Agencia Española de Protección de Datos (en adelante, AEPD) en su Informe sobre la *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción*, 2020, p. 5/53. Sin embargo, autores como BARRIO ANDRÉS, M., *op.cit.*, p. 57, diferencian la IA fuerte (a la que asimilan a la IA general y denominan también IA profunda) de la súper inteligencia. La idea fundamental es que estas inteligencias igualarían o incluso podrían llegar a superar a la

Ninguna de estas inteligencias artificiales se ha conseguido desarrollar todavía⁶⁴. El tipo de IA que ha disparado la aplicación práctica de esta disciplina es la que se conoce como *IA débil* y que, en contraste con las anteriores, se caracteriza por desarrollar soluciones capaces de resolver problemas o tareas concretas y acotadas. Estos sistemas, que dependen de un programador humano, carecen de capacidades cognitivas generales y no tienen conciencia ni comprensión del mundo que los rodea más allá de la tarea específica para la que han sido diseñados⁶⁵.

Por tanto, a pesar de la existencia actual de redes neuronales artificiales que se asemejan en gran medida a las redes neuronales humanas, por el momento no se han podido reproducir a través de estos sistemas ciertas capaci-

inteligencia humana. Define las *IA fuertes* como aquellas que tienen la capacidad de «razonar, representar el conocimiento, planificar, aprender, comunicarse en lenguaje natural e integrar todas estas habilidades hacia un objetivo común», y la *súper inteligencia* como la «capacidad de razonamiento y resolución de problemas notablemente superior a la de cualquier ser humano». Se asocia a la denominada «singularidad tecnológica», tras la cual no habrá diferencia entre humano y máquina. También puede verse con más detalle sobre otras clasificaciones de la IA, MUÑOZ VELA, J.M., *op.cit.*, pp. 92 a 97. Como explica SALAZAR GARCÍA, I., «Retos actuales de la ética en la inteligencia artificial», en *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, coord. por Marcelo Bauzá Reilly; Lorenzo Cotino Hueso (dir.), 2022, p. 57, hay expertos que auguran que estos sistemas inteligentes del futuro podrían crear otros algoritmos aún más inteligentes que ellos hasta llegar a la *singularidad tecnológica*, que vendría a ser el momento en que los robots dotados de IA superarían a los humanos en todos los aspectos y tomarían decisiones independientemente del criterio humano.

⁶⁴ Como señala MUÑOZ VELA, J.M., *op.cit.*, p. 95, la IA fuerte o general sería lo que todavía no existe porque los sistemas actuales no pueden «entender». En la *Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno*, adoptada por la Comisión Europea para la Eficacia de la Justicia (CEPEJ), en Estrasburgo, 3-4 de diciembre de 2018, también se afirma que las IA fuertes, que estarían equipados no solo con inteligencia sino también con conciencia, sigue siendo ficticias. Los sistemas de aprendizaje automático que se están desarrollando actualmente se describen como IA «débiles» y son capaces de extraer patrones complejos y aprender de grandes volúmenes de datos de manera eficiente y, a menudo, con altos niveles de precisión predictiva» (vid., n° 63).

⁶⁵ BARRIO ANDRÉS, M., *op.cit.* p. 57. Señala este autor que la IA débil sólo es adecuada para realizar la tarea para la que se ha diseñado el sistema y no tiene autoconciencia. Ejemplos de esta modalidad podrían ser el de Deep Blue de IBM que sólo juega al ajedrez, o asistentes como Alexa o Siri. Afirma también SALAZAR GARCÍA, I., en <https://www.noentiendonada.es/idoia-salazar-la-ia-puede-hacer-mucho-mal/> (último acceso el 29 de octubre de 2024), que estamos en un momento muy incipiente de la IA, en la llamada IA débil, donde los algoritmos de IA dependen de un programador humano y solo son capaces de hacer muy bien una tarea muy concreta. Otra cosa será en un futuro, si la IA sigue desarrollándose, cuando esta tecnología alcance (si es que llega a alcanzarla) el nivel fuerte. La IA fuerte o también llamada IA general, se refiere a la IA del futuro donde el comportamiento inteligente de la máquina estará mucho más desarrollado en tantas competencias como podría estarlo una persona, siendo este uno de los grandes miedos que despierta la tecnología, es decir, que los robots superen en inteligencia a los humanos. En ese momento, si los humanos no hemos sido capaces de acompañar de manera conveniente a esta tecnología, desde un punto de vista ético y sin permitir demasiada autonomía, podríamos tener un problema.

dades de la inteligencia humana como la autorreflexión, el sentido común o la comprensión de la causalidad. Podría decirse que la principal limitación de los sistemas inteligentes es que no pueden crear conocimiento por sí mismos, a pesar de que sus resultados sí que sirvan de base para ello.

Por último, contamos con la denominada *IA de Propósito General* (GPIA –*General Purpose Artificial Intelligence System*–), que se refiere a los sistemas de IA que tienen la capacidad de abordar una gama diversa de problemas en lugar de estar limitados a una tarea específica. Son sistemas avanzados y eficaces para realizar una serie de tareas distintas sin haber sido específicamente entrenados para cada una de ellas. Su grado de autonomía y habilidad viene determinado por varias características clave, como la capacidad de adaptarse o rendir bien en nuevas tareas que surjan en el futuro, la demostración de competencia en dominios para los que no ha sido entrenado de forma intencionada y específica, la capacidad de aprender a partir de datos limitados y el reconocimiento proactivo de sus propias limitaciones con el fin de mejorar su rendimiento.

Según el RIA, se trata de sistemas que no tienen un propósito inicial previsto, pero que pueden ser entrenados o modificados para cumplir un propósito que podría convertirlos en sistemas de alto riesgo. El art. 3.66 RIA los define como «*sistemas de IA basados en un modelo de IA de uso general y que pueden servir para diversos fines, tanto para su uso directo como para su integración en otros sistemas de IA*». A su vez, se entiende por «*modelo de IA de uso general*», «*un modelo de IA, también uno entrenado con un gran volumen de datos utilizando la autosupervisión a gran escala, que presenta un grado considerable de generalidad y es capaz de realizar de manera competente una gran variedad de tareas distintas, independientemente de la manera en que el modelo se introduzca en el mercado, y que puede integrarse en diversos sistemas o aplicaciones posteriores, excepto los modelos de IA que se utilizan para actividades de investigación, desarrollo o creación de prototipos antes de su comercialización*» (art. 3.63).

2.2. TOMA DE DECISIONES AUTOMATIZADAS E INTELIGENCIA ARTIFICIAL: ¿DISTINTAS DENOMINACIONES DE UN MISMO FENÓMENO?

La dificultad de definir la IA y la posible ambigüedad del término llevaron al Parlamento Europeo a proponer en su *Resolución de 20 de octubre de 2020 con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial (2020/2014(INL))* (en adelante, Resolución PE sobre RC de 2020), que se utilizara en su lugar la expresión «toma de decisiones automatizada» para hacer referencia a «*una situación en la que un usuario delega inicialmente una decisión, en su totalidad o en parte, en una entidad mediante*

un programa informático o un servicio. Esa entidad, a su vez, utiliza modelos de toma de decisiones ejecutados automáticamente para llevar a cabo una acción en nombre de un usuario o para informar la decisión del usuario a la hora de realizar una acción».

En el Derecho comparado, la *Algorithmic Accountability Act of 2022* (sección 2) emplea estos términos y dispone que: «Se entenderá por “*sistema automatizado de decisión*” cualquier sistema, programa informático o proceso (incluidos los derivados del aprendizaje automático, la estadística u otras técnicas de tratamiento de datos o inteligencia artificial y excluidas las infraestructuras informáticas pasivas) que utilice el cálculo, cuyo resultado sirva de base para una decisión o juicio».

A nuestro modo de ver, ambas expresiones no son totalmente equiparables. Por un lado, porque la IA es una denominación genérica que incluye varios enfoques y técnicas y que engloba, entre otras manifestaciones específicas, los mecanismos de toma de decisión automatizados⁶⁶. Por otro lado, porque no todos los sistemas de toma de decisiones automatizadas están basados en IA, sino que hay técnicas más sencillas que pueden sustentar estos sistemas como, por ejemplo, los procedimientos de análisis basados en reglas⁶⁷. Sobre esto, la Agencia Española de Protección de Datos (en adelante, AEPD) establece categóricamente que «no todo sistema que toma una decisión automatizada es IA, no toda IA es *Machine Learning*, ni todo lo que se publicita como IA es realmente IA [...]»⁶⁸.

A pesar de que en numerosos documentos de las instituciones europeas se habla del impacto que las *decisiones automatizadas* pueden tener en las personas, así como de las ventajas y riesgos que pueden implicar para los derechos fundamentales, no suelen definirlas⁶⁹. Tanto es así que el CESE,

⁶⁶ El *Informe sobre la inteligencia artificial en la era digital (2020/2266(INI))*, señala que «el término IA es un término genérico que abarca una amplia gama de tecnologías, técnicas y enfoques antiguos y nuevos que se entiende mejor bajo la denominación de «sistemas de inteligencia artificial», que se refiere a cualesquiera sistemas automatizados que a menudo apenas tienen nada más en común que estar guiados por un conjunto dado de objetivos definidos por el ser humano, con distintos grados de autonomía en sus acciones, e intervenir en predicciones, recomendaciones o tomas de decisiones basadas en los datos disponibles».

⁶⁷ ROIG, A., *Las garantías frente a las decisiones automatizadas: del Reglamento General de Protección de Datos a la gobernanza algorítmica*, Barcelona, 2020, p. 18.

⁶⁸ *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial...*, cit., p. 7.

⁶⁹ Por ejemplo, en la *Resolución del Parlamento Europeo, de 12 de febrero de 2020, sobre los procesos automatizados de toma de decisiones: garantizar la protección de los consumidores y la libre circulación de bienes y servicios* (2019/2915(RSP)), aps. 12 y 13., se dice que tales sistemas se basan en la recopilación de grandes volúmenes de datos y se subraya el impacto significativo que los mismos pueden tener en los consumidores, especialmente en los que se encuentran en situaciones de vulnerabilidad. También se insiste en la idea de que ha de ser posible solicitar la revisión y rectificación humanas de aquellas decisiones automatizadas que sean definitivas y permanentes. También en la *Resolución del Parlamento*

en su *Dictamen sobre Propuesta de Directiva para adaptar las normas de responsabilidad extracontractual en la Inteligencia Artificial*, recomendaba que se estableciera una definición jurídica de lo que constituye una «*decisión tomada por máquinas que utilizan IA*» a fin de acabar con la inseguridad jurídica existente⁷⁰.

Tampoco el RGPD recoge entre las definiciones del art. 4 en qué consisten, pero prohíbe con carácter general en su art. 22 las *decisiones individuales basadas únicamente en el tratamiento automatizado de datos*, incluida la elaboración de perfiles⁷¹, siempre y cuando produzcan efectos jurídicos o afecten significativamente de modo similar a los interesados. Sobre ellas se pronunció el GT29 en sus *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679*, estableciendo que las mismas «*representan la capacidad de tomar decisiones por medios tecnológicos sin la participación del ser humano*» y pueden basarse en cualquier tipo de datos.

Además, trata de deslindar las decisiones automatizadas de la elaboración de perfiles explicando que las decisiones automatizadas pueden llevarse a cabo con o sin elaboración de perfiles, y que la elaboración de perfiles puede darse sin realizar decisiones automatizadas, es decir, sin estar conectada a una toma de decisiones. No obstante, a efectos del presente trabajo, nos interesarán los perfilados en la medida en que sí lo estén, es decir, cuando sirvan de base para adoptar una decisión (por ejemplo, en base al perfil de riesgo realizado de forma automatizada sobre un cliente de un banco se le deniega un crédito). A pesar de que son actividades distintas, también aclara el GT29 que no tienen por qué ser necesariamente independientes en la medida en que «algo que empieza como un simple proceso de decisiones automatizadas

Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103(INL)) se pone de relieve la creciente utilización de los sistemas de toma de decisiones automatizadas y basadas en algoritmos, pero sin definirlos.

⁷⁰ Ap. 2.2. del *Dictamen del Comité Económico y Social Europeo sobre «Propuesta de Directiva del Parlamento Europeo y del Consejo relativa a la adaptación de las normas de responsabilidad civil extracontractual a la inteligencia artificial (Directiva sobre responsabilidad en materia de IA)»*, [COM(2022) 496 final - 2022/0303 (COD)](2023/C 140/05)

⁷¹ A diferencia de lo que ocurre con las decisiones automatizadas, el RGPD sí define la «elaboración de perfiles» en el ap. 4 del art. 4. Según dicho precepto, se trata de «*toda forma de tratamiento automatizado de datos personales consistente en utilizar datos personales para evaluar determinados aspectos personales de una persona física, en particular para analizar o predecir aspectos relativos al rendimiento profesional, situación económica, salud, preferencias personales, intereses, fiabilidad, comportamiento, ubicación o movimientos de dicha persona física*». Para NAVAS NAVARRO, S., «La perspectiva de género ...», cit., p. 14/22, los conceptos de tratamiento automatizado de datos y la elaboración de perfiles no significan lo mismo, ya que la elaboración de perfiles o «profiling» es un tipo de tratamiento automatizado que busca segmentar a los individuos situándoles en categorías y que se basa en la minería de datos.

puede convertirse en un proceso basado en la elaboración de perfiles, dependiendo del uso que se dé a los datos»⁷².

Sin perjuicio de lo afirmado, el Consejo de Europa, en su *Recomendación CM/Rec (2021)8 del Comité de Ministros a los Estados miembros en materia de protección de las personas con respecto al tratamiento automatizado de datos de carácter personal en el contexto de la elaboración de perfiles*, adoptada el 3 de noviembre de 2021, recoge una definición de estos sistemas, según la cual, un «sistema automatizado de toma de decisiones» es un sistema que utiliza el razonamiento automatizado para *ayudar o sustituir* a un proceso de toma de decisiones que, de otro modo, sería realizado por humanos (art. 1.1.h)⁷³.

También encontramos algunas propuestas o aproximaciones a este concepto en trabajos doctrinales. Para ESPIN ALBA, aunque la «toma de decisiones» es un concepto variable según los contextos de actuación, en el ámbito de la informática consistiría en un proceso donde las decisiones son tomadas por sistemas informáticos o algoritmos, en lugar de por seres humanos, y con este fin los sistemas de IA utilizan datos, reglas y modelos predefinidos para analizar información y generar recomendaciones o decisiones en numerosos contextos⁷⁴. Y siguiendo a ROIG, un «proceso de toma de decisión automatizado» sería aquel «proceso de base tecnológica que incluye un modelo de decisión, un algoritmo que permite la aplicación del modelo, unos datos usados por el algoritmo para aprender de los mismos o para analizarlos y un contexto político, jurídico y económico»⁷⁵.

⁷² La doctrina es unánime al considerar que la elaboración de perfiles y las decisiones individuales automatizadas son actividades distintas, aunque conectadas. Para NÚÑEZ SEOANE, J., «El derecho a la información y acceso al funcionamiento de los algoritmos que tratan datos personales», en *La regulación de los algoritmos*, coord. por Gustavo Manuel Díaz González y Alejandro José Huergo Lora (dir.) 2020, p. 302, no es fácil deslindar ambos conceptos en la medida en que suelen confluír en la práctica y su regulación tiende a solaparse en el RGPD. Explica muy bien las diferencias entre estas dos figuras relacionadas con los procesos de tratamiento de datos, LAZCOZ MORATINOS, G., «Análisis jurídico de la toma de decisiones algorítmica ...», *cit.*, p. 288, para quien la decisión automatizada tiene un ámbito de aplicación distinto y puede solaparse parcialmente con la elaboración de perfiles o derivar de ésta. Y ROIG mantiene que no se trata de ámbitos excluyentes o incompatibles, pues el RGPD contempla la posibilidad de conexiones entre ambas nociones, de forma que puede haber decisiones automatizadas con o sin elaboración de perfiles. Para una explicación más detallada, *vid.*, *op.cit.*, pp. 33 a 36.

⁷³ Disponible en <https://search.coe.int/cm/?i=0900001680a46148>.

⁷⁴ ESPÍN ALBA, I., *op.cit.*, pp. 150.

⁷⁵ ROIG, A., *op.cit.*, pp. 18-19, opta por considerar estas herramientas como un sistema, y no ya simples tecnologías. Con ello se persigue una aproximación global o holística del proceso de toma de decisión automatizado. Tanto dicho autor, como LAZCOZ MORATINOS, destacan que las decisiones basadas únicamente en el tratamiento automatizado de datos, ya sean éstos ofrecidos por las personas, observados o inferidos, suponen la posibilidad de tomar decisiones sin participación humana. *Vid.*, LAZCOZ MORATINOS, G., «Análisis jurídico de la toma de decisiones algorítmica...», *cit.*, p. 288.

A efectos de este trabajo nos centraremos en el uso de sistemas automatizados de toma de decisiones basados en tecnologías de IA. Particular interés presentan, además, los que se basan en aprendizaje automático puesto que esta rama de la IA es especialmente útil para dicha tarea. De hecho, la envergadura que ha adquirido a día de hoy el *machine learning*, y el *deep learning* o aprendizaje profundo como especialización del primero, se observa en el propio RIA que, como se dijo en el apartado anterior, *destaca que las técnicas que permiten a los sistemas llegar a inferencias incluyen, por un lado, estrategias de aprendizaje automático en virtud de las cuales los sistemas aprenden de los datos cómo alcanzar determinados objetivos, y por otro lado, estrategias basadas en la lógica y el conocimiento que infieren a partir de conocimientos codificados o de una representación simbólica de la tarea que debe resolverse*⁷⁶.

En suma, de la lectura de un buen número de documentos de las instituciones europeas puede extraerse que, en el contexto de la IA, por «sistema automatizado de decisión» se entiende aquel sistema basado en un proceso tecnológico que utiliza técnicas de IA para procesar datos y arrojar resultados en forma de predicciones, recomendaciones o decisiones con impacto en la vida de las personas y sin intervención humana directa.

No obstante, existen otros sistemas para la toma automatizada de decisiones que no se basan en IA pero que también pueden ocasionar daños susceptibles de reparación. Es por ello que también se tendrán en consideración en este trabajo, si bien partiendo de la base de que, en esos casos, a diferencia de los primeros, no se presenta tan compleja la determinación de la responsabilidad que dimana de su causación.

2.2.1. Decisiones plenamente automatizadas y decisiones semiautomatizadas

A mayor abundamiento, es preciso señalar que los sistemas que permiten la automatización de decisiones (conocidos también por sus siglas en inglés como *ADM Systems –Automated Decision Making o Automated Decision Systems–*),

⁷⁶ Vid., Considerando 12. Incluso en la Orientación General del Consejo de 25 de noviembre de 2022, dichas técnicas se contemplaban expresamente en la definición de los sistemas de IA. Así, el art. 3.1 establecía que por «sistema de inteligencia artificial» o «sistema de IA» había que entender «un sistema concebido para funcionar con elementos de autonomía que, a partir de datos e información generados por máquinas o por seres humanos, infiere la manera de alcanzar una serie de objetivos, utilizando para ello *estrategias de aprendizaje automático o estrategias basadas en la lógica y el conocimiento*, y produce información de salida generada por el sistema, como contenidos (sistemas de inteligencia artificial generativa), predicciones, recomendaciones o decisiones, que influyen en los entornos con los que interactúa el sistema de IA».

pueden ser de dos tipos en función de si el resultado del sistema sólo se utiliza como mero apoyo a la decisión que adoptará después el humano, o si el resultado se adopta y ejecuta de forma totalmente automática, es decir, sin intervención humana. El primer tipo de sistemas son parcialmente o semiautomatizados y se conocen como sistemas de apoyo a la decisión (*decision support systems*), mientras que el segundo grupo son plenamente automatizados y se conocen como sistemas expertos (*automated decision systems*)⁷⁷. De ahí que ROIG considere, con acierto, que estaremos ante una decisión automatizada cuando el proceso de decisión «se *delega* o se *apoya* en un sistema controlado por un algoritmo que suele partir de un análisis de datos»⁷⁸, ya que dicho concepto es lo suficientemente amplio para abarcar ambos tipos de sistemas.

Sin perjuicio de dicha distinción, existe un riesgo común a ambos modelos como es la adopción de decisiones sesgadas o discriminatorias⁷⁹. Téngase

⁷⁷ MUÑOZ GUTIÉRREZ, C., *op.cit.*, p. 278. La AEPD en su informe sobre *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial...*, *cit.*, p. 6/53, se refiere a esta cuestión cuando dice que la IA puede adoptar dos roles en la toma de decisiones: por un lado, el de ayuda para el proceso de decisión, proporcionando una inferencia o perfil sobre un sujeto o una situación, para que un ser humano tome la decisión final; y por otro lado, el de toma y ejecución de la decisión. Tal como explica ORWAT, C., *op.cit.*, pp. 21-22, en una decisión completamente automatizada, todos los pasos que llevan a la toma de la decisión los ejecuta el *software*. Los humanos establecen las reglas para la toma de decisiones o, en el caso de los métodos de aprendizaje automático, los algoritmos generan parte de esas reglas basándose en el análisis de datos. Ejemplos de sistemas totalmente automatizados podrían ser los casos en que un préstamo bancario se concede en línea y sin intervención humana, los sistemas de recomendación en el comercio electrónico, los ajustes de precios automatizados, o los filtros de spam en programas de correo electrónico. También puede verse, CITRON, D.K. «Technological due process», *Washington University Law Review*, vol. 85, n° 6, p. 1263, aunque la autora se refiere a estos sistemas como *plenamente automatizados* y *mixtos*. Respecto a estos últimos diferencia, a su vez, tres tipos de sistemas en función del grado o tipo de participación humana en las decisiones del sistema. Ejemplos de casos reales en los que la intervención humana impidió la calificación de la decisión algorítmica como plenamente automatizada, se recogen en BARROS VALE, S./ZANFIR-FORTUNA, G., *Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities*, Future of Privacy Forum, 2022, pp. 28-34. En todos ellos se entendió que la intervención humana en la evaluación de los resultados del algoritmo tenía un impacto significativo en la decisión final y que se habían tenido en cuenta para tomar las decisiones otros criterios o factores adicionales además de la decisión/recomendación del sistema.

⁷⁸ *Op.cit.*, pp. 18-19. No obstante, pone de relieve que actualmente la mayoría de las decisiones automatizadas no son totalmente autónomas o puramente automatizadas, sino más bien elaboración de perfiles más o menos orientados a sistemas de toma de decisiones. Sin embargo, en un futuro no muy lejano no será así porque muchas decisiones podrán formalizarse hasta el punto de ser totalmente autónomas, pese a la posibilidad de revisión periódica posterior —e incluso esta revisión podría automatizarse— (*vid.*, p. 35).

⁷⁹ Afirma CITRON, D.K. *op.cit.*, p. 1267, que independientemente de si el sistema es totalmente automatizado o mixto, los problemas que se generan son los mismos. Proporciona ejemplos de estos errores en pp. 1268 y ss, señalando que uno de ellos puede deberse a que los programadores se equivoquen al traducir a lenguaje informático las reglas o políticas sobre las que ha de adoptarse

en cuenta que cualquiera de ellos puede contener sesgos en el diseño o en los conjuntos de datos utilizados. Así, si la negativa para conceder un préstamo se ha basado en un perfilado previo hecho por el sistema de IA que ha tenido en cuenta alguna característica protegida de la persona (ej. género, raza, discapacidad, etc.), o la ha inferido de otros datos, a la que se ha asignado una ponderación negativa para valorar el nivel de riesgo del solicitante, será indiferente que la decisión final desfavorable sea automática (plenamente automatizada) o que, a la vista del resultado del algoritmo y confiando en ese criterio sin ponderar otros, una persona –humano en el circuito o *human in the loop*– decida denegarle el crédito (decisión semiautomatizada), ya que en ambos casos la decisión se habría basado en un perfilado discriminatorio.

Como en su momento se explicará con detalle⁸⁰, ha de tenerse en cuenta que el simple hecho de que intervenga un humano en la toma de decisión no garantiza que la misma deje de ser plenamente automatizada si éste se limita a aceptar y reproducir acríticamente los resultados del sistema. En la práctica, no es raro que esto ocurra debido a que las personas tendemos a confiar en exceso en la corrección de los resultados obtenidos por los programas informáticos creyendo, erróneamente, que son siempre neutrales y objetivos. A esto se le conoce como «sesgo de la automatización», es decir, la tendencia comprobada de los humanos a deferir sus decisiones a la máquina al considerar que el juicio del algoritmo es más acertado que el suyo propio cuando ambos no coinciden⁸¹. Se trata de un tipo de sesgo que, a pesar de estar presente en muchos ámbitos, supone un riesgo particular en aquellas personas que

una decisión dando lugar a decisiones injustas o ilícitas. Esto ocurrió en Colorado y California, entre otros casos que cuenta la autora, con sus sistemas de concesión de ayudas o subvenciones públicas que fueron un fracaso porque contenían reglas incorrectas y cometieron numerosos errores durante el tiempo que se utilizaron.

⁸⁰ En el apartado 2.5.2 del Capítulo III de esta obra.

⁸¹ Así lo expone EUBANKS, V., *Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor*, St Martin's Press, 2018, pp. 17-18 y 40-42. También trata esta cuestión CITRON, D.K. *op.cit.*, pp. 1271-1272, señalando que precisamente el sesgo de la automatización lleva a que, en la práctica, no haya tanta diferencia entre un sistema plenamente automatizado y uno «mixto» o semiautomatizado. Indica la autora que hay numerosos estudios que demuestran que los seres humanos confían en las decisiones automatizadas incluso cuando sospechan que el sistema funciona mal y que dicha confianza se debe a la percepción de que estos sistemas son más precisos y menos propensos a cometer errores que podrían cometer los humanos. Y también advierte ORWAT, C., *op.cit.*, p. 22, que incluso cuando se utilizan sistemas semiautomatizados, existe en la práctica una tendencia de los responsables humanos a asumir las decisiones del sistema. Esto ocurre por diferentes motivos, en particular, por razones de eficiencia, supuesta neutralidad o mayor objetividad de las decisiones algorítmicas, o porque les resulta difícil justificar a sus superiores por qué se apartan de la recomendación del sistema. Por todo ello, los sistemas de apoyo a la toma de decisiones tienden a convertirse, de hecho, en sistemas plenamente automatizados. Para RIVAS VALLEJO, P., «Sesgos de automatización...», *cit.*, p. 50, esto es así por el «poder retórico» que poseen los algoritmos en la medida en

utilizan *software* de apoyo para la toma de decisiones en análisis y diagnóstico (por ejemplo, médicos, gestores de riesgo o analistas financieros)⁸².

Como apunta SOLAR CAYÓN, la literatura científica en sociología cognitiva y economía conductual muestra que es psicológicamente muy difícil desatender las recomendaciones de los sistemas algorítmicos, por lo que es muy probable que las personas acaben siguiéndolas, especialmente, cuando se promueve el uso de estos sistemas bajo el argumento y con el objetivo de contribuir a erradicar la subjetividad y los errores humanos. Se pregunta asimismo el autor, con qué fundamento o legitimidad podríamos apartarnos de lo recomendado por el algoritmo si ni siquiera contamos con ciertos conocimientos técnicos o estadísticos para saber interpretar adecuadamente sus resultados⁸³.

Sobre este aspecto incidió el Tribunal Supremo de Wisconsin en su decisión sobre el caso *Loomis v. Wisconsin* donde tuvo en especial consideración lo manifestado por uno de los expertos que intervino como testigo, corroborando el gran desconocimiento que los tribunales tienen sobre estas herramientas y, en concreto, sobre las dirigidas a la evaluación de riesgos: «*El tribunal no conoce cómo COMPAS compara el historial del individuo con el del grupo de población con el que es comparado. El tribunal ni siquiera sabe si esa población es una población de Wisconsin, Nueva York, California. ... El tribunal no posee esta clase de información y lo que estamos haciendo es informar mal al tribunal cuando le proporcionamos estas gráficas y dejamos que las utilice para sentenciar*»⁸⁴. Precisamente por ello, dicho Tribunal advirtió de la necesidad de adoptar una serie de cautelas dirigidas a asegurar que los jueces no sobreestimasen el valor de los resultados del sistema en casos futuros⁸⁵, cautelas que a nuestro juicio deberían adoptarse no sólo en el ámbito judicial sino en cualquier otro en que se empleen sistemas de IA para la toma de decisiones.

que las decisiones que generan tienen una *poderosa legitimidad* e influyen en nuestras decisiones y las decisiones que otras personas toman sobre nosotros.

⁸² BELTRÁN DE HEREDIA, I., «Automatización y obsolescencia humana», en *Retos jurídicos de la inteligencia artificial*, coord. por Agustí Cerrillo i Martínez y Miquel Peguera Poch, 2020, pp. 119-120.

⁸³ SOLAR CAYÓN, J.I., *op.cit.*, p. 170.

⁸⁴ Extracto recogido por SOLAR CAYÓN, J.I., *op.cit.*

⁸⁵ El Tribunal de Wisconsin aclara, además, que el uso de estos sistemas por los jueces sólo debe ser válido cuando lo utilicen junto con otros factores o criterios para motivar sus sentencias, pero no como criterio exclusivo de decisión. Sobre este tema, añade HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, pp. 33-34, que en Cataluña la administración penitenciaria utiliza un programa parecido denominado *Riscanvi* para otorgar o denegar permisos penitenciarios y que la jurisprudencia considera, en sentido muy similar a la sentencia del Tribunal de Wisconsin, que se puede utilizar el resultado que arroja el programa para motivar la resolución (aunque no se conozca su contenido exacto), siempre que se puedan utilizar también otros factores para tomar la decisión final y someter a crítica el resultado que arroje el sistema, es decir, siempre que no se utilice de forma plenamente automatizada.

Consciente de los peligros que implica dicho sesgo a favor de la máquina, la Comisión Europea ha tratado de combatirlo con algunas de las previsiones recogidas en el RIA tendentes a conseguir que quienes se sirven de estos sistemas sean conscientes de cómo funcionan y puedan interpretar correctamente sus resultados⁸⁶. A este fin contribuyen, entre otros, el nuevo art. 4 que se ha incorporado en la versión definitiva del texto que lleva por título «Alfabetización en materia de inteligencia artificial». Dicho precepto dispone que *«Los proveedores y responsables del despliegue de sistemas de IA adoptarán medidas para garantizar que, en la mayor medida posible, su personal y demás personas que se encarguen en su nombre del funcionamiento y la utilización de sistemas de IA tengan un nivel suficiente de alfabetización en materia de IA, teniendo en cuenta sus conocimientos técnicos, su experiencia, su educación y su formación, así como el contexto previsto de uso de los sistemas de IA y las personas o los grupos de personas en que se utilizarán dichos sistemas»*.

También se mantiene el contenido previsto en el anterior art. 9.4.c (que en la versión final pasa a ser el art. 9.5.c) que establece que, a la hora de determinar las medidas de gestión de riesgos más adecuadas para estos sistemas, se procurará *«proporcionar la información requerida conforme al artículo 13 y, cuando proceda, impartir formación a los responsables del despliegue. Con vistas a eliminar o reducir los riesgos asociados a la utilización del sistema de IA de alto riesgo, se tendrán debidamente en cuenta los conocimientos técnicos, la experiencia, la educación y la formación que se espera que posea el responsable del despliegue, así como el contexto en el que está previsto que se utilice el sistema»*.

Además, los deberes de transparencia y comunicación de información que debe proporcionarse a los finalmente denominados «responsables del despliegue» (*usuarios* en las versiones anteriores) establecidos en el art. 13, también van dirigidos a que estos puedan interpretar correctamente los resultados arrojados por el sistema. Así, su primer apartado impone la obligación de diseñar y desarrollar los sistemas de IA con un nivel de transparencia que les permita *interpretar y usar correctamente su información de salida*. Además, en la versión definitiva del art.13.3.b), se ha añadido un *apartado vii)* que establece que la información de las instrucciones de uso que deben facilitarse a dichos sujetos deberá contener *«en su caso, información que permita a los responsables del despliegue interpretar la información de salida del sistema de IA de alto riesgo y utilizarla adecuadamente»*.

⁸⁶ También considera muy preocupante este sesgo, aunque no lo denomina así, la AEPD en su *Informe sobre la Adecuación del RGPD a la IA*, cit., p.7, donde señala que el sesgo humano a la hora de interpretar los resultados de la IA *«consiste en aceptar, sin espíritu crítico, los resultados de una IA como ciertos e inamovibles, asumiendo un “principio de autoridad” derivado de las expectativas creadas por dichos sistemas»*.

También dispone el art. 14.4 RIA (art. 14.3 en la versión anterior⁸⁷) que *las medidas de vigilancia o supervisión humana de los sistemas de alto riesgo deben ir dirigidas a permitir que las personas físicas a quienes se encomiende dicha vigilancia puedan, en función de las circunstancias y de manera proporcionada a estas: a) entender adecuadamente las capacidades y limitaciones pertinentes del sistema de IA de alto riesgo y poder vigilar debidamente su funcionamiento, por ejemplo, con vistas a detectar y resolver anomalías, problemas de funcionamiento y comportamientos inesperados; b) ser conscientes de la posible tendencia a confiar automáticamente o en exceso en la información de salida generada por un sistema de IA de alto riesgo («sesgo de automatización»), en particular con aquellos sistemas que se utilizan para aportar información o recomendaciones con el fin de que personas físicas adopten una decisión; c) interpretar correctamente la información de salida del sistema de IA de alto riesgo, teniendo en cuenta, por ejemplo, los métodos y herramientas de interpretación disponibles; d) decidir, en cualquier situación concreta, no utilizar el sistema de IA de alto riesgo o desestimar, invalidar o revertir la información de salida que este genere; e) intervenir en el funcionamiento del sistema de IA de alto riesgo o interrumpir el sistema pulsando un botón de parada o mediante un procedimiento similar que permita que el sistema se detenga de forma segura».*

En definitiva, para evitar que las personas encargadas de valorar los resultados del sistema los asuman de forma acrítica, se torna imprescindible que conozcan cómo funcionan y cuáles son sus principales características, ya que sólo así estarán en disposición de poder interpretarlos correctamente y de tenerlos en consideración (o no) en su justa medida. Si carecen de una formación mínima en este campo, no podrán ser conscientes, por ejemplo, de que los resultados pueden discriminar a personas o colectivos vulnerables, o de que estos sistemas no se basan en certezas sino en probabilidades generales pudiendo llegar a inferencias erróneas sobre personas concretas⁸⁸. Con esa intervención humana en la toma de decisiones basadas en los resultados del sistema se podrá garantizar que las particularidades de cada sujeto sean

⁸⁷ Dispone el art. 14.3 que estas medidas de vigilancia serán definidas por el proveedor, bien para integrarlas en el sistema cuando sea técnicamente viable, bien para que las adopte el propio responsable del despliegue del sistema antes de su introducción en el mercado o puesta en servicio, pero sin concretar la forma en que en que se debería articular este mecanismo de salvaguarda. Como critica SORIANO ARNANZ, A., «La Propuesta de Reglamento de Inteligencia Artificial de la UE y los Sistemas de Alto Riesgo», *Revista General de Derecho de los Sectores Regulados*, n.8, noviembre 2021, p. 19, «este precepto da libertad a los proveedores para decidir la forma en que se articulará este mecanismo, no estableciendo si quiera qué se considera una suficiente intervención humana. Es decir, por una parte, recoge de manera mucho más completa que el RGPD la forma en que se podría articular la supervisión humana y las cuestiones que se podrían tener en cuenta al establecer este mecanismo, pero no fija ningún mandato específico más allá de la obligación genérica de intervención humana».

⁸⁸ Todas estas características de los sistemas de *machine learning* se analizarán *infra*, en el apartado 2.3.3.

analizadas y tenidas en cuenta para adoptar la decisión que le afecte y que se respete, en definitiva, lo que se ha venido a denominar el «debido proceso tecnológico»⁸⁹.

Por último, hay que diferenciar las decisiones automatizadas que afectan directamente a personas físicas de aquellas que no lo hacen. Ejemplos de las primeras podrían ser, entre otros muchos, los sistemas de autenticación de usuarios, de evaluación crediticia o de selección de personal, y ejemplos de las segundas, los sistemas de control industrial, optimización de la logística y transporte de mercancías o gestión de inventarios. Además, como indica la AEPD, en caso de que se tomen decisiones que afectan a las personas, las mismas pueden ser relativas a la interacción de la persona en su contexto social, como el acceso a un contrato o servicio, o relativas a la personalización de dicho servicio, como podría ser la personalización en los mandos de un coche o la programación de un televisor. Las decisiones pueden hacer predicciones sobre la evolución del sujeto, realizar una evaluación sobre el estado actual de éste, o bien decidir la ejecución de un conjunto de acciones⁹⁰. En cualquier caso, las decisiones automatizadas que interesan en este trabajo son aquellas que afectan directamente a personas o grupos, en concreto, por sus efectos discriminatorios.

Si bien no es momento de profundizar en el tema (aunque sí lo haremos más adelante) debe advertirse ya de entrada que, como regla general, la toma de decisiones *individualizadas* basadas únicamente en el tratamiento de *datos personales* que tengan *efectos jurídicos o similares* en los interesados está prohibida por el art. 22 RGPD. No obstante, dicha regla presenta importantes excepciones basadas en el propio consentimiento del interesado o en que sean necesarias para la celebración de un contrato⁹¹.

⁸⁹ En este sentido, SORIANO ARNANZ, A., «La Propuesta de Reglamento de Inteligencia Artificial...», *cit.*, pp. 18-19; SOLAR CAYÓN, J.I., *op.cit.*, p. 153, define el «*Debido proceso tecnológico*» como el conjunto de procedimientos que garanticen que los algoritmos predictivos satisfacen ciertos estándares de evaluación y revisión al objeto de asegurar su imparcialidad y corrección. También sobre el «debido proceso tecnológico», puede verse CITRON, D.K., *op.cit.* De su trabajo puede extraerse que, en la era de la automatización de las decisiones administrativas, hay que garantizar que se respetan los derechos procesales con el fin de asegurar que las decisiones automatizadas sean transparentes, revisables y estén sujetas a un control humano significativo. La autora propone procedimientos como la posibilidad de auditar los algoritmos, mecanismos de revisión judicial y diseñar los sistemas de forma que los ciudadanos puedan comprender las decisiones que les afectan y cuestionarlas.

⁹⁰ Así lo afirma en su informe sobre *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial...*, *cit.*, p. 6/53, poniendo como ejemplo un sistema decisonal que, en el ámbito sanitario, pudiera prever la evolución de un paciente, realizar un diagnóstico automático tras su exploración o prescribir un determinado tratamiento.

⁹¹ Vid., art. 22 RGPD.

El hecho de que el precepto citado y los derechos asociados a las decisiones automatizadas que contempla el RGPD se refieran únicamente a las decisiones *plenamente* automatizadas y no a las semiautomatizadas (a pesar de que ambos tipos de sistemas podrían arrojar resultados discriminatorios), supone una importante traba en la lucha contra la discriminación algorítmica⁹². Por ello, debe plantearse la ampliación de su ámbito de aplicación también a los casos en que las decisiones sean semiautomatizadas. Así lo proponía ya en 2018 el GT29 en sus *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679*⁹³ y parece ser la postura acogida en la reciente sentencia del TJUE de 7 de diciembre de 2023, sobre la que volveremos en el capítulo III donde se tratará con detalle el art. 22 RGPD.

2.3. *MACHINE LEARNING: CONCEPTO, TIPOS, ELEMENTOS Y CARACTERÍSTICAS*

2.3.1. Concepto y tipos de *machine learning*

Especialmente relevante para la toma de decisiones se erige la rama de la IA conocida como *machine learning* (en adelante, ML) o aprendizaje automático. Se trata de un género dentro de la IA débil que se caracteriza por desarrollar soluciones capaces de resolver un problema concreto y acotado, funcionando el *software* como si fuera humanamente inteligente para tareas concretas⁹⁴.

⁹² GINÈS i FABRELLAS, A., *op.cit.*, p. 319.

⁹³ Vid., p. 28.

⁹⁴ Una de las definiciones más influyentes y referenciadas en trabajos académicos sobre IA y aprendizaje automático, es la proporcionada por Tom Mitchell en su obra de 1997, titulada *Machine Learning*. Lo define de la siguiente manera: «Un programa de ordenador aprende de la experiencia *E* con respecto a una tarea *T* y una medida de rendimiento *P*, si su rendimiento en la tarea *T*, medido por *P*, mejora con la experiencia *E*». Esta definición se considera fundamental porque enmarca el ML como un proceso basado en la mejora del rendimiento a través de la experiencia acumulada, lo que refleja la esencia de los algoritmos de aprendizaje automático que ajustan su comportamiento a partir de los datos observados. No obstante, una definición menos técnica podría ser la proporcionada por la Royal Society en *Machine learning: the power and promise of computers that learn by example. An introduction*, 2017, recuperada de <https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-introduction.pdf> (última consulta el 22 de octubre de 2024), donde se afirma que se trata de una forma de IA que permite a las máquinas aprender a partir de ejemplos, datos y experiencias en lugar de seguir instrucciones preprogramadas, lo que significa que los sistemas pueden mejorar su rendimiento en tareas específicas analizando grandes volúmenes de datos. Por su parte, la AEPD en *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial...*, cit., p. 6/53, define el ML como la rama de la IA que «diseña modelos predictivos que construyen por sí mismos la relación entre las variables a estudiar mediante el análisis de un conjunto inicial de datos, la identificación de patrones y el establecimiento de criterios de clasificación. Una vez fijados los criterios, al introducir

Esta técnica describe una amplia gama de algoritmos que proporcionan a los sistemas la capacidad de aprender y mejorar automáticamente a partir de la experiencia y en base a grandes volúmenes de datos sin estar programados expresamente, esto es, sin necesidad de basarse en reglas previamente codificadas⁹⁵. Así, lo que atribuye singularidad al *machine learning* respecto de los sistemas basados en instrucciones previamente programadas o codificadas (sistemas o programas conocidos como *deterministas*) es precisamente esta capacidad de aprender con un cierto grado de autonomía (mayor o menor en función del tipo de algoritmo utilizado) y en base normalmente a un entrenamiento previo a través de grandes cantidades de datos.

Por tanto, mientras en la programación tradicional no había aprendizaje de ningún tipo, sino que el programa hacía lo que se le ordenaba de forma axiomática sin sondear más datos y el código no cambiaba por su cuenta, los programas que utilizan IA son diferentes. Se les entrena con enormes conjuntos de datos y a partir de esos datos de entrada el algoritmo genera unas reglas que van ligadas a la probabilidad con la que ocurren los hechos en esa base de datos. De ahí que el algoritmo pueda extraer reglas del tipo «*si raza=negro; entonces préstamo= denegado con probabilidad del 80 %*», o «*si raza = negro y número de hijos>4 entonces préstamo=aceptado con probabilidad del 3 %*). En este tipo de sistemas, que simulan la inteligencia humana imitando comportamientos como la toma de decisiones o la resolución de problemas, sí hay aprendizaje porque dependiendo de los datos de entrada el algoritmo genera unas reglas u otras, pudiendo cambiar esas reglas automáticamente mientras está funcionando. Es decir, el modelo de inferencia se ajusta dinámicamente en función de cada

un nuevo conjunto de datos el componente IA es capaz de realizar una inferencia». Y en la *Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno*, p. 25, se dice que «El aprendizaje automático es un campo de la informática en el que los programas informáticos aprenden de la experiencia. Los algoritmos hacen que una máquina realice un proceso de entrenamiento, como un niño que aprende en su entorno».

⁹⁵ Así lo define, desde el ámbito jurídico, BARRIOS ANDRÉS, M., *op.cit.*, p. 59. También en la doctrina jurídica, LAZCOZ MORATINOS, G., «Análisis jurídico de la toma de decisiones algorítmica...», *cit.*, p. 284, señala que a grandes rasgos se trata de complejos modelos algorítmicos que, frente a la programación clásica, tienen la capacidad de aprender en su tarea programada, rutinaria y automatizada, a partir de ejemplos que se le facilitan y además con la capacidad de manejar un número de variables inabordable por la programación clásica. Y MUÑOZ VELA, J.M., *op.cit.*, p. 87, lo define, por un lado, «como el método de aprendizaje de la IA que toma como *input* datos y genera como *output* un modelo que puede ser usado para resolver un problema o para la toma de decisiones. El proceso de generación del modelo entrenado con datos sería el aprendizaje; y por otro lado, como «conjunto de algoritmos que proporcionan a los sistemas la capacidad de aprender y mejorar automáticamente a partir de la experiencia y acumulación de datos sin necesidad de estas programados previamente o de tener reglas previamente codificadas. Un ejemplo de ello sería el software de reconocimiento de imágenes. Este método de aprendizaje identifica patrones en conjuntos de datos objeto de análisis, construye modelos y hace predicciones sin disponer de reglas o modelos predefinidos».

nuevo conjunto de datos de entrada, refinando las relaciones ya establecidas. Por ello se dicen que son sistemas *adaptativos*⁹⁶.

Dentro del aprendizaje automático se pueden distinguir, a su vez, varios tipos (o clases de algoritmos) que varían en complejidad y autonomía: el basado en reglas, el supervisado, no supervisado, de refuerzo y sus variantes, que emplean distintas técnicas⁹⁷. A grandes rasgos, y sin ánimo de ser excesivamente técnicos, se puede mantener que mientras que los sistemas de aprendizaje basados en reglas o instrucciones explícitamente determinadas no pueden ser entrenados y cada mejora necesita ser programada, el resto de modalidades mencionadas aprende a base de ser entrenado y alimentado con nuevos datos de entrada. Por ello, cuando el sistema encuentra una correlación significa que acaba de hallar un fenómeno que desconocía hasta el momento y esto implica hallar nueva información que se reintroduce en los algoritmos para que realicen búsquedas más exactas, de modo que los resultados anteriores perfeccionan el funcionamiento del algoritmo y el sistema aprende⁹⁸. No obstante, como apunta BARRIO ANDRÉS, actualmente los sistemas basados en reglas también pueden incluir capacidades de *machine learning* dando lugar a sistemas híbridos⁹⁹.

⁹⁶ Así describe los sistemas adaptativos la AEPD en su Informe sobre la *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial...*, cit., p. 6/53. Y ORWAT, C., *op.cit.*, p. 22 diferencia entre los «sistemas estáticos», que son aquellos que realizan el análisis de datos una sola vez o en intervalos separados de tiempo y adaptan las reglas para tomar la decisión, y los «sistemas dinámicos», que adaptan y optimizan constantemente las reglas o modelos de decisión mediante el análisis continuo de flujos de datos. Como apunta NAVAS NAVARRO, S., «‘Deshumanizando’ al contrato: una mirada desde la inteligencia artificial», en *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, coord. por Natalia Álvarez Lata; José Manuel Busto Lago (pr.), 2024, p. 36, hemos pasado de la lógica formal basada en reglas del tipo «si ...entonces» (*rule based knowledge*) para inferir el conocimiento, a la extracción del mismo a partir de datos que se presentan de forma desestructurada (*data-learned knowledge*).

⁹⁷ Explica brevemente cada uno de ellos el Grupo de Expertos de Alto Nivel sobre IA de la Comisión Europea en el documento «Una definición de la Inteligencia Artificial: principales capacidades y disciplinas científicas», 2019, p.4. En la doctrina jurídica, HUERGO LORA, A., «Una aproximación a los algoritmos desde el derecho administrativo», en A. Huergo Lora (dir.) y G. Díaz González (coord.), *La regulación de los algoritmos*, Navarra, 2020, p. 37, explica que «se habla de aprendizaje supervisado (*supervised learning*) en el caso de los antiguos programas que jugaban al ajedrez (a los que no sólo se les programaban las reglas del juego sino que se les indicaba cuáles eran las mejores jugadas ante los distintos escenarios posibles, frente al “aprendizaje no supervisado” (*unsupervised learning*) de los actuales algoritmos predictivos, porque (tomando una vez más el ejemplo de “AlphaGo”) es “el ordenador” el que, aplicando el algoritmo, consigue, a partir del análisis de millones de partidas, identificar las mejores jugadas, es decir, aquellas que ofrecen más probabilidades de ganar. De aquí la expresión *machine learning* (aprendizaje mecánico, es decir, “no humano”), porque, una vez puesto en marcha el mecanismo, el sistema “aprende solo”, sin necesidad de guía».

⁹⁸ Así lo explica EGUILUZ CASTAÑEIRA, J.U., «Desafíos y retos que plantean las decisiones automatizadas y los perfiles para los derechos fundamentales», *Estudios de Deusto: revista de Derecho Público*, Vol. 68, N.º. 2, 2020, p. 331.

⁹⁹ BARRIO ANDRÉS, M., *op.cit.*, p. 63.

Asimismo, encontramos especializaciones del ML como el *deep learning* (DL) o aprendizaje profundo¹⁰⁰, cuya principal base tecnológica son las redes neuronales que son capaces de identificar patrones ocultos y crear correlaciones inesperadas entre *inputs* y *outputs* simplemente a partir del análisis de datos masivos, generando automáticamente los modelos predictivos o las reglas de decisión que posteriormente aplican a nuevos conjuntos de datos¹⁰¹.

Tal como señala NAVAS NAVARRO, de los diferentes tipos de aprendizaje de un sistema de IA, la Propuesta de RIA se refería en un primer momento solo al basado en aprendizaje supervisado como demostraba su art. 10, sin contemplar en su ámbito de aplicación los sistemas de IA cuyo aprendizaje era no supervisado. Pero dicha situación cambia a raíz del surgimiento de los denominados modelos fundacionales, como ChatGPT, que fueron incorporados en la versión de la Propuesta de 14 de junio de 2023 publicada con las enmiendas del Parlamento¹⁰². Dichos sistemas consisten en «sistemas avanza-

¹⁰⁰ La expresión *deep learning* (aprendizaje profundo) se utiliza con el mismo sentido, y hace referencia, como el *machine learning*, a que el sistema aprende autónomamente, a partir del análisis de gran cantidad de datos con un modelo matemático de gran complejidad (es decir, profundo). La técnica está inspirada en el funcionamiento básico de las neuronas del cerebro y aunque existe hace más de 50 años, ahora tenemos suficiente volumen de datos y capacidad de computación para aplicarla en una multitud de casos. La principal diferencia entre el DL y el ML es la estructura de la arquitectura de red neuronal subyacente, ya que, mientras los modelos tradicionales de ML utilizan redes neuronales simples (con una o dos capas computacionales), los de DL utilizan tres o más capas (normalmente cientos o miles de capas para entrenar los modelos). Sobre esta técnica señala el Grupo de Expertos de Alto Nivel sobre IA de la Comisión Europea en su documento «Una definición de la Inteligencia Artificial: principales capacidades y disciplinas científicas», 2019, p. 4 que es más precisa que el ML y requiere un menor grado de orientación humana.

¹⁰¹ Así lo explica, SOLAR CAYÓN, J.I., *op.cit.*, p. 148. También apunta BARRIOS ANDRÉS, M., *op.cit.*, p. 62, que el término «profundo» hace referencia al gran número de capas de neuronas ocultas en la red. A diferencia del aprendizaje automático supervisado, en el que han de realizarse manualmente los procesos iniciales de extracción de características relevantes y diseño de un modelo capaz de categorizar o clasificar la información que se desea analizar, en los de aprendizaje profundo estos procesos de extracción de características y modelización son automáticos. Los hace el sistema a partir de imágenes, textos o sonidos. Son altamente precisos y en ocasiones superan el rendimiento humano. Y señala SÁNCHEZ CAPARRÓS, M., *op.cit.*, p. 4/34, que este tipo de sistema, aunque es más eficiente para lograr resultados, también «representa un enorme desafío para los derechos humanos, ya que para que la IA sea fiable es esencial comprender por qué un sistema se ha comportado de una determinada manera y por qué ha ofrecido una respuesta específica, y éste es justamente el talón de Aquiles de una técnica que, por su propia naturaleza, conlleva que no sea posible explicar los millones de correlaciones “...que se procesan en las capas ocultas de la red (...) en el sentido de que se pueda ofrecer una explicación detallada de lo que ocurrió allí...”».

¹⁰² Indica esta autora que en los casos de aprendizaje supervisado «son los informáticos, ingenieros, y otros expertos, los que etiquetan los datos, los depuran, eliminan datos erróneos o corruptos, intentan eliminar en la medida de lo posible los sesgos que presenten, determinan qué datos son lo más representativos posibles, completos; en definitiva, que poseen la calidad necesaria para que los resultados que produzca el sistema sean lo más precisos posibles y las decisiones que, en su caso,

dos de *machine learning* que, una vez han aprendido patrones y correlaciones a partir de los datos, aunque no se les haya dicho explícitamente que lo hagan, son capaces de generar nuevos casos, situaciones o ejemplos a partir de los datos de entrenamiento. Estos modelos necesitan una gran cantidad de datos, lo que conlleva que el sistema deba entrenarse no solo con datos predeterminados o con datos sintéticos, sino que también necesitan entrenarse con datos de libre acceso en internet y, además, que lo haga constantemente de acuerdo con un modelo de aprendizaje autosupervisado por el propio sistema [...].».

Por otro lado, en función del grado de predeterminación de los resultados que genera el sistema, debe quedar claro también que mientras aquéllos basados en instrucciones condicionales del tipo «*si X, entonces Y*», carecen de autonomía decisional en la medida en que el sistema genera de forma automática un resultado que se corresponde con las instrucciones predeterminadas por el programador (por ejemplo, *si se cumplen los requisitos fijados en la norma para ser beneficiario del bono social de electricidad, entonces concesión de la ayuda*), en los de aprendizaje automático –en sus distintas variantes y especializaciones–, los resultados no están totalmente predeterminados desde el momento en que son diseñados, sino que van actualizándose y perfeccionándose conforme van analizando nuevas y mayores cantidades de datos¹⁰³.

Si bien con cualquiera de estos sistemas se pueden cometer errores o se puede llegar a resultados sesgados o decisiones discriminatorias que ocasionen daños a personas concretas (piénsese que han podido introducirse instrucciones o variables incorrectas en el sistema, puede haberse configurado mal la variable objetivo, etc.), el riesgo que implican para los derechos fundamentales los sistemas con autonomía decisional y los problemas para determinar los elementos o presupuestos clásicos de la responsabilidad civil se acrecientan cuando se utilizan algoritmos de *machine learning* –y principalmente de *deep learning*–, ya que su opacidad y complejidad técnica impiden controlar su funcionamiento de manera efectiva y descubrir cómo o por qué han arrojado un determinado resultado dañoso y quien habría de responder del mismo¹⁰⁴.

tome sean también lo más acertadas posibles». Vid., NAVAS NAVARRO, S., «'Deshumanizando' al contrato...», *cit.*, pp. 36-37.

¹⁰³ SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 88, señala que en este tipo de sistemas los resultados no vienen predeterminados por las personas encargadas de diseñarlos, sino que estas se limitan a alimentar los programas con datos y a proporcionar una serie de instrucciones más o menos extensas, pero que, *en todo caso, dejan un margen de apreciación al propio sistema*. Por tanto, los sistemas de autonomía decisional valorarán el supuesto de hecho concreto en función de lo que han ido aprendiendo e irán ajustando los resultados a medida que cambie el contexto y que obtengan *feedback* sobre las decisiones previamente adoptadas.

¹⁰⁴ Profundizaremos sobre esta cuestión en el capítulo V.

Así lo explica la Comisión Europea en su *Informe sobre las repercusiones en materia de seguridad y responsabilidad civil de la IA, el internet de las cosas y la robótica*, de 19 de febrero de 2020, donde dice que el funcionamiento autónomo de una IA implica la ejecución de una tarea sin que cada paso esté predefinido, y que dicha ejecución se lleva a cabo con menor control o, en última instancia, sin ningún control o supervisión humanos inmediatos. Por ello, los algoritmos basados en el aprendizaje automático de la máquina pueden ser difíciles, si no imposibles, de comprender («efecto caja negra») ¹⁰⁵. Además, como señala el Parlamento Europeo, esta técnica es capaz de convertir datos aparentemente inocuos en datos sensibles y personales con las consiguientes vulneraciones del derecho fundamental a la protección de datos personales ¹⁰⁶.

Es indudable, por tanto, que son los sistemas de IA que tienen cierta autonomía o margen de apreciación a la hora de tomar decisiones los que exigen mayor atención, en comparación con los sistemas clásicos que serían aquellos que frente al supuesto de hecho establecen un resultado predeterminado por el programador ¹⁰⁷. En estos últimos, a diferencia de los primeros, no parece tan complicado determinar la responsabilidad por daños. Sin embargo, los algoritmos avanzados de ML generan problemas éticos y jurídicos, sobre todo a efectos de determinar la responsabilidad civil por los daños que puedan llegar a causar.

Aun con todo, es importante no perder de vista que, como también apunta la Comisión en el *Libro Blanco sobre la inteligencia artificial* ¹⁰⁸, incluso en los casos en que estos sistemas funcionen de manera autónoma a partir de su percepción del entorno y sin seguir un conjunto predefinido de instrucciones, «su comportamiento lo definen y restringen en gran medida sus desarrolladores. Los objetivos los definen las personas, y los sistemas de IA deben optimizarse para alcanzarlos». Y en sentido similar se pronuncia el Consejo de Europa en su *Recomendación del Comité de Ministros a los Estados miembros en materia de protección de las personas con respecto al tratamiento automatizado de datos*

¹⁰⁵ COM (2020) 64 final, p. 17/20.

¹⁰⁶ Vid., el estudio del Parlamento Europeo titulado «*The ethics of artificial intelligence: Issues and initiatives*», p. 13, donde se presentan ejemplos de lo afirmado en el texto. Recuperado de [https://www.europarl.europa.eu/thinktank/es/document/EPRS_STU\(2020\)634452](https://www.europarl.europa.eu/thinktank/es/document/EPRS_STU(2020)634452)

¹⁰⁷ El RIA acoge esta distinción y para que el sistema caiga bajo su ámbito de aplicación, esto es, para que se considere un sistema de IA a los efectos del Reglamento debe ser un sistema «*diseñado para funcionar con distintos niveles de autonomía*» (art. 3.1). Por tanto, la idea de la autonomía es fundamental e implica, como explica con claridad HUERGO LORA, en <https://almacenederecho.org/por-que-aciertan-las-sentencias-sobre-el-algoritmo-del-bono-social-electrico>, que el sistema no se limita a realizar operaciones a partir de reglas dadas y de los datos de entrada como si fuera una calculadora, sino que «*pone algo de su parte*», incluso con «*capacidad de adaptación tras el despliegue*».

¹⁰⁸ COMO (2020) 65 final, p. 20/31.

de carácter personal en el contexto de la elaboración de perfiles, al afirmar que «todos los sistemas automatizados de toma de decisiones están diseñados por humanos y tienen cierto grado de participación humana en su funcionamiento. Los responsables en última instancia de cómo un sistema recibe sus entradas (por ejemplo, quién recoge los datos que se alimentan a un sistema), cómo se utiliza y cómo se interpretan sus resultados y se actúa en consecuencia son personas»¹⁰⁹.

Queda claro, por tanto, que a la hora de imputar responsabilidades debe tenerse muy presente que en el diseño y despliegue de estos sistemas siempre hay intervención humana en mayor o menor medida.

2.3.2. Elementos del *machine learning*: algoritmos y datos

2.3.2.1. Algoritmos

A pesar de la complejidad que entraña concretar en qué consiste la IA en general o el *machine learning* en particular, de lo que no cabe duda es de que dicha tecnología se nutre de dos componentes indispensables para funcionar: los algoritmos y los datos¹¹⁰. De hecho, el gran crecimiento que ha experimentado la IA en los últimos años se vincula con el acelerado desarrollo que se ha experimentado para obtener ambos elementos.

Un algoritmo puede definirse como un conjunto ordenado y finito de operaciones que permite hallar la solución a un problema a partir de la entrada de información (datos/*inputs*)¹¹¹. En sentido figurado, es común entre los

¹⁰⁹ CM/Rec (2021)8. Vid., ap. 2.4, que finaliza diciendo que «Los sistemas (especialmente los basados en IA) deben permitir la intervención humana operativa siempre que sea apropiado o necesario para garantizar su funcionamiento legítimo, incluso en lo que respecta a los principios de equidad y no discriminación».

¹¹⁰ Así lo explicaba la Comisión Europea en su *Libro Blanco sobre IA* donde señala que los algoritmos son entrenados para *inferir determinados modelos a partir de un conjunto de datos*, a fin de determinar las acciones que se requieren para alcanzar un objetivo determinado.

¹¹¹ Así lo define la primera acepción del término «algoritmo» del D.R.A.E. Prácticamente idénticas a la ofrecida por la RAE son las definiciones acogidas por las instituciones y doctrina. Por ejemplo, en la *Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno* del Consejo de Europa, se define como «secuencia finita de reglas formales (operaciones o instrucciones lógicas) que hacen posible obtener un resultado a partir de la entrada de información», secuencia que puede ser parte de un proceso automatizado de ejecución y que se basa en modelos diseñados a través de aprendizaje automático. Entre los autores, PEÑA MARÍ, R., *De Euclides a Java, la historia de los algoritmos y de los lenguajes de programación*, 2006, lo define como el «conjunto de reglas que, aplicada sistemáticamente a unos datos de entrada apropiados, resuelven un problema en un número finito de pasos elementales». De la definición se infiere, como indica RIVAS VALLEJO, P.,

autores equiparar el algoritmo con una receta de cocina en la que los ingredientes serían los datos. Y de la misma forma que ocurre con los ingredientes a la hora de cocinar, cuanto de mayor calidad sean los datos de que se nutre al sistema, mejor será el resultado (decisión algorítmica) que se obtenga¹¹².

Los algoritmos pueden implementarse o no a través de programas informáticos. En este último caso, huelga mencionar la importancia que ostentan en el ámbito de la ingeniería informática, ya que permiten al programador resolver el problema para el que busca solución antes de escribirlo en un lenguaje de programación. De hecho, los programas informáticos no son otra cosa que una serie compleja de algoritmos escritos en un lenguaje de programación que pueden ser ejecutados en un ordenador. De ahí la importancia de los algoritmos informáticos en el diseño, creación y desarrollo de la IA¹¹³.

A efectos del presente trabajo interesan los algoritmos que permiten encontrar soluciones a un problema de *forma automatizada*. Además, de entre todos los tipos de algoritmos que existen y que pueden cumplir funciones muy diversas, los que en la práctica vienen funcionando como sistemas de IA son los *algoritmos predictivos*, es decir, aquellos que combinando múltiples datos de entrada a través de técnicas estadísticas y lógicas, permiten determinar, analizar y predecir hechos o ciertos aspectos de la personalidad o el comportamiento, intereses y hábitos de una persona, en base a los cuales se toman decisiones¹¹⁴. Actualmente son relevantes en la medida en que *generan o sustentan* decisiones que pueden ir desde la fijación de precios de un producto en función de la estimación (predicción) que hace el sistema de cuánto estarían dispuestos a pagar los consumidores por un producto o servicio en virtud de

«Análisis desde el derecho antidiscriminatorio», en P. Rivas Vallejo (dir.) *Discriminación algorítmica en el ámbito laboral*, Navarra 2022, p. 439, que hay dos elementos cruciales desde una perspectiva jurídica: la secuencia de instrucciones y la información o datos que utiliza, sobre los que, a su vez, habrá que analizar jurídicamente su propiedad y derechos de uso, según su procedencia, y el posible acceso a ellos por cualquier impugnante. Estos temas serán tratados más adelante en este trabajo.

¹¹² Entre otros, ESPINOZA VALENCIA, F.F., «Ese algoritmo, el discriminador», recuperado de <https://www.abogacia.es/publicaciones/blogs/blog-de-innovacion-legal/ese-algoritmo-el-discriminador/> (última consulta: 2 de abril de 2024); y CASTILLA, K., *Cuatro ángulos de análisis de la igualdad y la no discriminación en la inteligencia artificial*, Barcelona, 2022, p. 20, accesible en <https://www.idhc.org/arxius/recerca/Inteligencia-artificial-vF.pdf>

¹¹³ CASTILLA, K., *op.cit.*, p. 12.

¹¹⁴ HUERGO LORA, A., «Administraciones Públicas e inteligencia artificial: ¿más o menos discrecionalidad?», *El Cronista del Estado Social y Democrático de Derecho*, n° 96-97 (Octubre-Noviembre), 2021 (Ejemplar dedicado a: La discrecionalidad de los reguladores), p. 3/20, accesible en <https://laadministracionaldia.inap.es/noticia.asp?id=1512494>. En sentido similar, GINÉS i FABRELLAS, A., *op.cit.*, p. 298, considera que la base de los sistemas de IA es hacer predicciones de conducta.

De hecho, en el *Informe sobre la inteligencia artificial en la era digital*, se recuerda que la IA se basa en programas informáticos que utilizan modelos probabilísticos y algoritmos predictivos para un conjunto de objetivos específicos.

sus hábitos de consumo, hasta el diagnóstico médico de una enfermedad en función de los síntomas y otros datos de las personas.

Otros ejemplos de este tipo de algoritmos utilizados con frecuencia en la práctica son los utilizados en el sector bancario para evaluar la solvencia crediticia de los clientes y «predecir» las probabilidades de impago, los empleados en el sistema judicial para calcular las probabilidades de reincidencia de los presos o por la Administración para determinar la probabilidad de que los contribuyentes defrauden al fisco, así como los que utilizan los empleadores para la selección y cribado de currículums de candidatos a un puesto de trabajo para encontrar la persona «probabilísticamente más idónea» para el mismo.

En relación a los sistemas de IA utilizados para valorar la concesión de permisos penitenciarios, explica HUERGO LORA que «[e]n un modelo de predicción algorítmica, se fija un *objetivo* (evitar la reincidencia o la comisión de delitos durante el disfrute del permiso) y el modelo analiza los datos de todos los permisos concedidos, para tratar de encontrar qué factores diferenciales existen entre los permisos que acaban con la comisión de un delito y aquellos otros que no dan problemas. De ese modo, el programa aísla unos factores (a los que se dota de un peso o ponderación) que son los que se toman en consideración para conceder o denegar el permiso»¹¹⁵. Y también relata con claridad el funcionamiento de las predicciones basadas en datos para valorar la concesión o denegación de préstamos. De forma resumida, podría decirse que estos modelos tienen en cuenta numerosas variables relacionadas con el objetivo que se quiere conseguir (en este caso, la devolución del préstamo). Una vez combinados todos los factores que ayudan a predecir si el crédito será devuelto o no (en este ejemplo concreto la mayoría estarán correlacionados con la solvencia del peticionario) *a los que el propio algoritmo atribuye un peso o importancia concreta*, el sistema arroja una fórmula que se puede posteriormente aplicar a quien solicita un préstamo. De esta forma, cuando después se introduzcan los datos personales de cada solicitante (entre ellos, titulación académica, nivel de ingresos, historial

¹¹⁵ HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, p. 47, nota al pie 38. Como explica este autor, no se trata de un «baremo» al uso establecido de forma transparente por un órgano administrativo porque no es una norma la que fija los criterios a tener en cuenta para valorar la concesión o no del permiso, así como el peso de cada uno de ellos, sino que estos han sido deducidos por el algoritmo a partir de experiencias (datos) anteriores y podrán ir cambiando en virtud de los nuevos datos que se vayan suministrando. Es decir, nadie fija expresa y conscientemente los factores del baremo, sino que éste lo crea el propio sistema informático cuando, tras analizar los datos introducidos sobre el fenómeno en cuestión, encuentra correlaciones. Sobre el uso de estos algoritmos para la toma de decisiones concluye el autor en la p. 85, que si bien la utilización de algoritmos no es más arriesgada que la utilización de baremos en general y de hecho elimina algunos de sus problemas, es necesario evitar que el algoritmo tenga en cuenta factores vedados por el artículo 14 CE, y es preciso poder demostrar que no se han tenido en cuenta.

asegurador y crediticio, etc.), el resultado del algoritmo indicará las *probabilidades* de que devuelva el préstamo, y ese dato podrá ser tenido en cuenta (o no) por el encargado de tomar la decisión (en este caso, el Banco) para conceder o denegar el crédito y/o el tipo de interés aplicable¹¹⁶.

2.3.2.2. Datos

Los datos, entendidos en sentido amplio, son el conjunto de informaciones que procesan o requieren los algoritmos para ejecutar las instrucciones que se les dan y buscar los resultados que se espera obtener¹¹⁷.

Si bien desde antaño se ha dicho que «la información es poder», hoy en día esta idea está más en boga que nunca. De hecho, se viene afirmando que los datos son el «cuarto poder»¹¹⁸ y se consideran «el petróleo» del siglo XXI o incluso «el nuevo oro»¹¹⁹. Según establece la Agenda España Digital 2025, se estima que el volumen de datos generados en el mundo en el año 2025 será de 175 zettabytes¹²⁰.

¹¹⁶ HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, p. 30 y ss, y concretamente, p. 32.

¹¹⁷ Según la primera acepción del D.R.A.E, se trata de «información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho»; y en su tercera acepción (para el contexto de la informática), sería la «información dispuesta de manera adecuada para su tratamiento por una computadora». Sobre dicho concepto, puede verse ampliamente, LÓPEZ-TARRUELLA MARTÍNEZ, A., *Propiedad intelectual e innovación basada en los datos*, Navarra, 2021, pp. 34 y ss, quien diferencia entre *datos*, *información*, *conocimiento* y *sabiduría*, señalando lo siguiente: «Los “datos” son unidades básicas de información recolectadas, pero que están desorganizadas o sin procesar. Estos datos devienen “información” cuando son analizados y se identifican relaciones entre ellos. El “conocimiento” se genera por analistas que reconocen la importancia de la información en un determinado contexto (se puede poseer mucha información, pero no extraer valor, si no se es capaz de detectar cual de ella es importante para adoptar una decisión); y, por último, la “sabiduría” se genera a partir de las decisiones que se toman con base en los datos analizados. En esta jerarquía (datos, información, conocimiento y sabiduría), cada etapa posterior es dependiente de la anterior: no hay sabiduría sin conocimiento, no hay conocimiento sin información, no hay información sin datos». Igual que hace este autor en la obra citada, optaremos por manejar un concepto amplio de «dato» que no solo abarca los datos en sentido estricto sino también la información (que, a su vez, puede constituir en conocimiento o sabiduría) representados en formato digital.

¹¹⁸ De «cuarto poder» y «nuevo poder social» habla RIVAS VALLEJO, P., «Sesgos de automatización...», *cit.*, p. 50 y doctrina citada.

¹¹⁹ Así opina V. MAYER-SCHÖNBERGER, profesor de regulación y gestión de internet en el Internet Institut de la Universidad de Oxford, y uno de los expertos más reconocidos internacionalmente en el mundo de los datos masivos. Vid. https://www.eldiario.es/tecnologia/diario-turing/big-data_1_5767121.html

¹²⁰ MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL, *España Digital 2025*, p. 67. Téngase en cuenta la siguiente equivalencia: 1 Zettabyte (ZB) = 1000 Exabytes = 1 millón de petabytes = 1000 millones de terabytes.

Como ya hemos apuntado, la IA se sirve de ingentes cantidades de datos para su funcionamiento y posterior perfeccionamiento porque, en principio, cuantos más datos se utilicen –siempre y cuando sean datos «de calidad»–, mejores pueden ser los resultados que ofrezca el sistema¹²¹. Ese volumen masivo de datos, al que también se hace referencia con el término *macrodatos* (en inglés, *Big Data*), es un pilar fundamental de la IA y se basa en la extracción y procesamiento de grandes cantidades de datos a elevada velocidad¹²².

El Grupo de trabajo del artículo 29 (en adelante, GT29)¹²³ propuso una definición de *Big Data*, estableciendo que consistía en la creciente capacidad de la tecnología para respaldar la recopilación y el almacenamiento de grandes cantidades de datos, así como para analizar, comprender y extraer el valor total de los datos¹²⁴. Tiempo después, la *Resolución de 14 de marzo de 2017 del Parlamento Europeo sobre las implicaciones de los macrodatos en los derechos fundamentales* establecía con algo más de detalle que «El concepto de macrodato se refiere a la recopilación, análisis y acumulación constante de grandes cantidades de datos, incluidos personales, procedentes de diferentes fuentes y objeto de tratamiento automatizado mediante algoritmos informáticos y avanzadas técnicas de tratamiento de datos, utilizando tanto datos almacenados como datos transmitidos en flujo continuo, con el fin de generar correlaciones, tendencias y patrones (analítica de macrodatos)». Sobre esta base, creemos que podría definirse el *Big Data* a efectos de este trabajo como el conjunto de recursos o tecnologías cuyo objetivo consiste en tratar ingentes cantidades de información empleando complejos algoritmos y estadísticas con una pluralidad de finalidades, destinadas a facilitar la toma de decisiones¹²⁵.

¹²¹ No obstante, apunta RIVAS VALLEJO, P., «Sesgos de automatización...», *cit.*, p. 55, que un exceso de datos puede llegar a provocar más correlaciones o patrones falsos de los sistemas.

¹²² En el *Libro Blanco sobre IA*, la Comisión Europea califica a los datos como la «piedra angular de la IA». Señala ALCOLEA AZCÁRRAGA, C., «La responsabilidad patrimonial de la Administración y el uso de algoritmos», *Revista General de Derecho Administrativo*, n.º 59, enero 2022 (recuperado de <https://laadministracionaldia.inap.es/noticia.asp?id=1512629>) que el término «*Big Data* o *macrodatos* expresa la cantidad de información que es posible contener y manejar, de un tamaño imposible para los humanos. Al hablar de este, se suele hacer referencia a las 'V': volumen, variedad, velocidad, valor y veracidad».

¹²³ Se trata de un grupo de trabajo europeo independiente que se ha ocupado de cuestiones relacionadas con la protección de la privacidad y los datos personales hasta el 25 de mayo de 2018 (entrada en vigor del RGPD). Fue el predecesor del actual Comité Europeo de Protección de Datos (CEPD) cuyo objetivo es garantizar que la normativa de protección de datos se aplique de manera coherente en los países de la UE, así como en Noruega, Liechtenstein e Islandia. Se puede consultar información sobre el trabajo de dicho grupo en: https://edpb.europa.eu/about-edpb/more-about-edpb/article-29-working-party_es.

¹²⁴ Vid., Comisión Europea (GT29). *Opinion 03/2013 on Purpose Limitation*. 2013, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf, p. 45 (Última consulta: 3 de abril de 2024).

¹²⁵ GARRIGA DOMÍNGUEZ, A., *Nuevos retos para la protección de datos personales en la Era del Big Data y de la computación ubicua*. Madrid, 2015, p. 28. No obstante, como recalcan MAYER-

Huelga mencionar que, en la actualidad, los datos pueden provenir de las fuentes más diversas (por ejemplo, sistemas de reconocimiento biométrico, redes sociales, historiales y búsquedas de Internet, aplicaciones informáticas, aparatos de seguimiento, etc.) y que su extracción y tratamiento pueden poner en serio peligro derechos fundamentales como la intimidad, la protección de datos personales o la igualdad y no discriminación.

Además, los datos no tienen valor en sí mismos, sino que han de estar estructurados para ser aprovechables. De hecho, expertos afirman que la única manera de explotar esos volúmenes masivos de datos es a través de sistemas de IA que permitan procesarlos de manera eficiente y ayuden (a los humanos) a tomar decisiones acertadas en base a ellos¹²⁶. Tal como explica MERCADER UGUINA «los algoritmos se alimentan de datos y las posibilidades de análisis de macrodatos (*Big Data*), la IA y el aprendizaje automático permiten reelaborar gran cantidad de informaciones simples, de forma que, combinadas entre sí a través de técnicas estadísticas y lógicas, es posible determinar, analizar, predecir ciertos aspectos de la personalidad o el comportamiento, los intereses y los hábitos de una persona». Y para GIL GONZÁLEZ, más que su volumen o naturaleza, lo importante es el valor potencial de los datos que puede explotarse a través de tecnología especializada con el fin de descubrir o inferir hechos y tendencias ocultos en las bases de datos¹²⁷. Ese es el objetivo principal del *Big Data*, esto es, aportar, descubrir o comprender cosas que cuando se analizaban pequeñas cantidades de información no podían conocerse.

El procesamiento de los *macrodatos* ha adquirido dimensiones colosales a las que se ha hecho referencia con múltiples términos como *dataísmo*, *capitalismo cognitivo*, *sociedad del cálculo* o *datificación*¹²⁸. Respecto a este último término,

SCHÖNBERGER, V./ CUKIER, K., *op.cit.*, p.17, «no existe ninguna definición rigurosa de los datos masivos. En un principio, la idea era que el volumen de información había aumentado tanto que la que se examinaba ya no cabía en la memoria que los ordenadores emplean para procesarla, por lo que los ingenieros necesitaban modernizar las herramientas para poder analizarla. Ese es el origen de las nuevas tecnologías de procesamiento, como Map-Reduce, de Google, y su equivalente de código abierto, Hadoop, que surgió de Yahoo. Con ellos se pueden manejar cantidades de datos mucho mayores que antes, y esos datos –esto es lo importante– no precisan ser dispuestos en filas ordenadas ni en las clásicas tabulaciones de una base de datos». Sin perjuicio de lo dicho, el D.R.A.E. también recoge una definición de «macrodatos» como el conjunto de datos que, por su gran volumen, requieren técnicas especiales de procesamiento.

¹²⁶ Así lo entiende SALAZAR GARCÍA, I., en <https://www.noentiendonada.es/idoia-salazar-la-ia-puede-hacer-mucho-mal/> (último acceso el 29 de octubre de 2024).

¹²⁷ GIL GONZÁLEZ, E., *Big Data, Privacidad Y Protección De Datos*, Accedit 2015 al Premio de Protección de Datos Personales de Investigación de la Agencia Española de Protección de Datos, 2016, p. 18.

¹²⁸ A todos ellos hace mención RIVAS VALLEJO, P., «Sesgos de automatización...», *cit.*, p. 50-51.

«datificar» consiste en procesar información hasta convertirla en datos. Tal como explican MAYER-SCHÖNBERGER Y CUKIER, «Datificar se refiere a recopilar información sobre cuanto existe bajo el sol –incluyendo cosas que en modo alguno solíamos considerar información antes, como la localización de una persona, las vibraciones de un motor o la tensión que soporta un puente–, y transformarla a formato de datos para cuantificarla. Esto nos permite usar la información de modos nuevos, como en el análisis predictivo: detectar que un motor es proclive a un fallo mecánico basándonos en el calor o en las vibraciones que emite. Lo que se consigue así es liberar el valor latente e implícito de la información»¹²⁹.

Los conjuntos de datos (*datasets*) que necesitan los sistemas de IA para funcionar los recoge y define recientemente el art. 3 RIA. Básicamente serían los datos de entrenamiento, los de validación y los de prueba. Los primeros son los «usados para entrenar un sistema de IA mediante el ajuste de sus parámetros entrenables» (art.3.29); los segundos, los «usados para proporcionar una evaluación del sistema de IA entrenado y adaptar sus parámetros no entrenables y su proceso de aprendizaje para, entre otras cosas, evitar el ajuste insuficiente o el sobreajuste» (art. 3.30)¹³⁰; y los terceros, «los datos usados para proporcionar una evaluación independiente del sistema de IA, con el fin de confirmar el funcionamiento previsto de dicho sistema antes de su introducción en el mercado o su puesta en servicio» (art.3.32). Además, se contemplan otros conjuntos de datos como los «datos de entrada» que se definen como los «proporcionados a un sistema de IA u obtenidos directamente por él a partir de los cuales produce la información de salida» (art. 3.33), los «datos personales» y «no personales», así como lo que se entiende por «categorías especiales de datos personales», que interesan especialmente en este

¹²⁹ MAYER-SCHÖNBERGER, V./ CUKIER, K., *Big data, la revolución de los datos masivos*, 2013, en la p. 28. Y en la p. 95 dicen que «Datificar un fenómeno es plasmarlo en un formato cuantificado para que pueda ser tabulado y analizado» y destacan que es algo muy diferente a la «digitalización». También entre nuestros autores, LÓPEZ-TARRUELLA MARTÍNEZ, A., *op.cit.*, p. 23, señala que el «concepto de digitalización ha dejado paso al de “datificación”, término creado a partir de la combinación de los vocablos datos y cuantificación, y que hace referencia a la transformación del mundo en datos procesables y cuantificable». Y según afirma GIL GONZÁLEZ, E., *op.cit.loc.cit.*, «el *big data* permite transformar en información muchos aspectos de la vida que antes no se podían cuantificar o estudiar, como los datos no estructurados (por ejemplo, datos no-texto como fotografías, imágenes y ficheros de audio) [...] Así, nuestra localización ha sido dataficada, primero con la invención de la longitud y la latitud, y en la actualidad con los sistemas de GPS controlados por satélite. Del mismo modo, nuestras palabras ahora son datos analizados por ordenadores mediante minería de datos. E incluso nuestras amistades y gustos son transformados en datos, a través de los gráficos de relaciones de redes sociales o los “likes” de Facebook».

¹³⁰ También se definen en el apartado 31) los «conjunto de datos de validación» como un conjunto de datos independiente o una parte del conjunto de datos de entrenamiento, obtenida mediante una división fija o variable.

trabajo y que se definen por remisión al RGPD (vid., apartados 50, 51 y 37 respectivamente)¹³¹.

2.3.3. Características de los sistemas de *machine learning*

Los sistemas de ML tienen una serie de características que deben tenerse en consideración para comprender la diversa problemática jurídica que generan:

En primer lugar, los algoritmos de ML se basan en matemáticas, estadísticas y probabilidades, pero no en certezas¹³². Es decir, no son más que aplicaciones que calculan la probabilidad (general) de que una persona (concreta) pueda ser un estudiante ejemplar, un prestatario de riesgo, un mal trabajador o un potencial maltratador, a partir del procesamiento de grandes volúmenes de datos¹³³. Precisamente por esto pueden adoptarse decisiones erróneas, ya que lo que es verdad para el grupo, no tiene por qué serlo necesariamente para una persona concreta¹³⁴. En este sentido, la *Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno*¹³⁵ señala que «la singularidad de los sistemas actuales de procesamiento de *big data* es que no intentan reproducir nuestro modelo cognitivo sino que producen estadísticas de contexto sobre un tamaño de datos sin precedentes, *sin ninguna garantía real de excluir correlaciones falsas*»¹³⁶. De hecho, siempre hay un porcentaje de falsos positivos o falsos negativos en los resultados que estos sistemas arrojan.

¹³¹ También se contempla un tipo de datos personales como los «biométricos» (ap. 34) y los «datos operativos sensibles» (ap. 38). Una clasificación jurídica de los datos a efectos de su tratamiento informático puede verse en LÓPEZ-TARRUELLA MARTÍNEZ, A., *op.cit.*, pp. 44-58.

¹³² Sobre el carácter estadístico de los resultados del algoritmo cobra particular relevancia el pronunciamiento del Tribunal Supremo de Wisconsin en su decisión sobre el conocido caso del algoritmo COMPAS, comentado entre otros por SOLAR CAYÓN, J.I., *op.cit.loc.cit.*

¹³³ Explica LAZCOZ MORATINOS, G., «Análisis jurídico de la toma de decisiones algorítmica...», *cit.*, p. 287, que la aplicación de estas tecnologías sobre nuestros datos da lugar a inferencias y predicciones sobre nuestros comportamientos, preferencias, estado de salud, etc., conocidos como «datos inferenciales».

¹³⁴ SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 93. Advierte esta autora del riesgo que estos sistemas suponen para la singularidad, autonomía e intimidad de los individuos.

¹³⁵ Adoptada por la Comisión Europea para la Eficacia de la Justicia (CEPEJ), en Estrasburgo, 3-4 de diciembre de 2018.

¹³⁶ P. 26; también los autores han incidido en que las predicciones de estos sistemas no está libres de errores puesto que se basan en modelos matemáticos que puede estar mal diseñados y dependen de unos datos que pueden no predecir adecuadamente el futuro porque se hayan producido cambios de uno u otro tipo. Así, HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, p. 32

Dicha característica ha conllevado que se haya criticado esta forma de adoptar decisiones porque, a pesar de que estos sistemas son capaces de crear perfiles muy precisos, suponen un riesgo para la singularidad de los individuos y pueden vulnerar el derecho constitucional al debido proceso legal que incluye el derecho a una sentencia individualizada¹³⁷. También se ha afirmado que, en la medida en que comportamientos pasados de personas *similares* al interesado permiten predecir (con mayor o menor probabilidad) la conducta futura de dicho interesado, son incompatibles con las ideas de dignidad de la persona y de libre desarrollo de la personalidad del art. 10.1 CE, y, precisamente por ello, debería rechazarse que las decisiones con efectos directos sobre personas concretas y basadas en su conducta pudieran tomarse en función de predicciones algorítmicas¹³⁸.

En segundo lugar, a pesar de que son múltiples las funciones que puede tener un algoritmo de ML (entre ellas, segmentar clientes creando perfiles, ofrecer nuevos productos y servicios a medida del consumidor, etc)¹³⁹, son particularmente útiles para identificar patrones o correlaciones entre los datos pasados que analiza (es decir, datos referidos a situaciones u operaciones anteriores cuyo resultado se conoce) y de cuyo análisis infiere, automáticamente, las características que llevaron a los mejores resultados y toma estos criterios como base para adoptar decisiones futuras¹⁴⁰. No obstante, esto no significa que comprendan la causalidad, es decir, la relación causa-efecto que

¹³⁷ Así lo expone SOLAR CAYÓN, J.I., *op.cit.*, p. 168.

¹³⁸ Así, HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, p. 79, quien además sustenta su postura en que las normas actuales no lo permiten (salvo en casos excepcionales, añadimos nosotros), ni tendría sentido que fuesen modificadas para admitir esa posibilidad, puesto que esas predicciones conllevan, por un lado, un alto riesgo de llegar a inferencias erróneas, y por otro lado, pueden vulnerar el art. 10 CE. Sin embargo, considera perfectamente válido y posible que la Administración utilice predicciones algorítmicas, junto con otros factores, para ejercer potestades discrecionales en las que la norma no vincula el contenido de la decisión administrativa a la constatación de uno o varios hechos, sino que establece un marco dentro del que es válida cualquier decisión que tome la Administración siempre que esté adecuadamente motivada y se hayan respetado las normas procedimentales. El problema, para este autor, es «el paso de las decisiones administrativas de alcance amplio (muchas veces reglamentarias y/ o de planificación), que inevitablemente se basan en juicios probabilísticos o prospectivos (susceptibles de error, tanto si se basan en algoritmos como si no lo hacen), a decisiones que tienen efectos directos sobre personas concretas y que se basan en su conducta».

¹³⁹ Aplicando distintas técnicas analíticas e instrumentos se les entrena para una u otra función (por ejemplo, a través de técnicas de agrupamiento un algoritmo puede aprender a identificar puntos en común o correlaciones entre grandes cantidades de datos aparentemente muy diversos; o a través de técnicas de regresión se les puede entrenar para el cálculo de probabilidades). Vid., GERARDS, J./XENIDIS, R., *op.cit.*, p. 33.

¹⁴⁰ HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, p. 35. En el mismo sentido, MUÑOZ VELA, J.M., *op. cit.*, p. 87.

puede haber, o no, entre esos datos. «Correlación» y «causalidad» no son conceptos equivalentes en la materia que tratamos. Por ello, que exista correlación entre variables, no siempre significa que una variable sea la que causa la otra; pero el algoritmo identifica correlaciones porque es mucho más sencillo correlacionar que detectar causas¹⁴¹. Entender esto es fundamental porque la forma tradicional de adoptar decisiones, basada en criterios racionales (es decir, en algún principio o razonamiento) o irracionales (criterios subjetivos como la sana crítica del experto en la materia) ha cambiado en el escenario algorítmico, donde los criterios anteriores se sustituyen por predicciones basadas en correlaciones identificadas analizando operaciones anteriores (datos pasados)¹⁴². No deja de ser cuanto menos llamativo que mediante el análisis de datos pasados se alcancen predicciones futuras.

El problema reside, como explica SORIANO ARNANZ, en que actualmente estos sistemas se consideran tan fiables que sus resultados, a pesar de basarse en correlaciones, se están tratando como si fueran relaciones causales aunque no exista prueba de la causalidad. Esto genera el riesgo de que se produzcan inferencias erróneas, absurdas o ilógicas (por ejemplo, que llevar calcetines de ejecutivo provoca infartos o que el consumo de margarina aumenta el número de divorcios)¹⁴³. Es evidente, por tanto, que se están adoptando deci-

¹⁴¹ RIVAS VALLEJO, P., «Sesgos de automatización...», *cit.*, p. 48. Proporciona SÁNCHEZ CAPARRÓS, M., *op.cit.*, p.5/34, un ejemplo muy ilustrativo sobre esta idea: «yo puedo observar que cuando canta el gallo de mi vecina amanece –hay aquí una correlación–. No obstante, eso no significa que el canto del gallo de mi vecina sea lo que provoca –o causa– el amanecer. Es decir, no hay relación de causalidad entre el canto del gallo y el amanecer, más allá de que pueda correlacionar ambos hechos porque muy a menudo se presentan juntos».

¹⁴² Sobre este cambio radical en la forma de actuar y tomar decisiones alertaban ya hace una década MAYER-SCHÖNBERGER, V./ CUKIER, K., *op.cit.*, p. 18. Afirmaban en su obra que «la era de los datos masivos pone en cuestión la forma en que vivimos e interactuamos con el mundo. Y aun más, la sociedad tendrá que desprenderse de parte de su obsesión por la causalidad a cambio de meras correlaciones: ya no sabremos por qué, sino solo qué. Esto da al traste con las prácticas establecidas durante siglos y choca con nuestra comprensión más elemental acerca de cómo tomar decisiones y aprehender la realidad». En esta misma línea, considera ALCOLEA AZCÁRRAGA, C., *op.cit.*, que «la pretensión racionalizadora tradicional del Derecho puede encontrarse con una crisis en sus bases puesto que en muchos ámbitos tradicionales del Derecho el concepto de causalidad es clave, mientras que la IA y el *Big Data* funcionan bajo la premisa de la correlación».

¹⁴³ A esto se le conoce como correlaciones espurias, es decir, relaciones estadísticas entre dos o más variables que no resultan de ninguna conexión causal directa, sino que son resultado de una coincidencia o de una tercera variable no considerada. También se habla de «alucinaciones de los algoritmos». Ejemplos de este tipo de correlaciones (o casualidades, que no «causalidades») pueden verse en <https://www.xataka.com/magnet/a-margarina-divorcios-11-divertidos-ejemplos-que-correlacion-no-implica-causalidad>. SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 91. Como sigue diciendo esta autora, es probable que exista una correlación entre las compras de freidoras realizadas por internet y los malos hábitos alimenticios. Sin embargo, puede ser que, en un caso concreto, una persona compre una freidora como regalo para otra. Si esta informa-

siones equivocadas basadas en factores ajenos al riesgo real e irrelevantes con respecto al hecho que se pretende proteger.

Por lo que se refiere a los sesgos discriminatorios, el problema acaece cuando el algoritmo tiene en cuenta características o variables protegidas (como la raza, el sexo o la discapacidad, entre otras) y las correlaciona directa o indirectamente con resultados negativos¹⁴⁴. Por ejemplo, cuando ser mujer se correlaciona con un peor desempeño profesional, no debido a una relación causal, sino a que históricamente las mujeres han sido evaluadas peor que los hombres por hacer el mismo trabajo.

En tercer lugar, no interpretan el sentido de las cosas, no pueden poner las cosas en contexto, porque su contexto es muy reducido y se limita al conjunto de datos con los que se entrena al sistema para que aprenda. De aquí se derivan dos ideas importantes: por un lado, la importancia de los datos de alimentación, ya que al ser ésta su única fuente de información y conocimiento, si los datos aportados al sistema no son de calidad porque son incompletos, incorrectos, sesgados o poco representativos, los resultados serán erróneos, parciales, sesgados o discriminatorios¹⁴⁵; por otro lado, que la imparcialidad estrictamente matemática, es por sí misma insuficiente para garantizar la imparcialidad desde una perspectiva política, jurídica o moral. Por ello, es necesario que los resultados estadísticos de estos sistemas se pongan en relación con el contexto social, institucional y jurídico en el que el sistema opera¹⁴⁶.

En cuarto lugar, son vulnerables por confusión, es decir, se puede engañar a la IA para confundir una tortuga con un rifle o para que el piloto automático de un vehículo autónomo se equivoque y no respete una señal de *stop*¹⁴⁷.

Por último, una de las características que mayores problemas plantea es la opacidad, tanto técnica como jurídica, de estos sistemas. De ella se dará am-

ción llega a los algoritmos empleados por su seguro médico, que desconocerán (o no serán capaces de comprender la “causa” —el propósito— con el cual se ha comprado la freidora, esa correlación podría contribuir a aumentar la prima del seguro de esa persona.

¹⁴⁴ GINÈS i FABRELLAS, A., *op.cit.*, p. 310.

¹⁴⁵ Esto se conoce como *GIGO*, del inglés *Garbage In, Garbage Out*, lo que viene a significar que si al sistema de IA le entra basura (datos de mala calidad), sus resultados también serán basura. Para FERNÁNDEZ DE LA MORENA, B., *Discriminación Algorítmica. Estudio del sesgo en arquitecturas de aprendizaje profundo*, Universidad Autónoma de Madrid, 2019, pp. 1-2, se puede considerar que, en cierto modo, un algoritmo es tan bueno (o tan malo, añadimos nosotros) como la información con la que trabaja.

¹⁴⁶ SOLAR CAYÓN, J.I., *op.cit.*, p. 165.

¹⁴⁷ Sobre estos inconvenientes puede verse PARRA, V., «Inteligencia artificial para la toma de decisiones», en <https://www.mytips.es/inteligencia-artificial-para-la-toma-de-decisiones/>. También pueden consultarse las siguientes noticias: <https://www.kaspersky.es/blog/rsa2021-tesla-mobileye-perception-gap/25378/>; <https://www.xataka.com/vehiculos/enganan-a-tesla-pase-56-a-140-km-h-automaticamente-pegando-trozo-cinta-a-senal>

plia cuenta en el capítulo III. Baste explicar por ahora respecto a la primera que, en ocasiones, en función de la complejidad del algoritmo empleado (por ejemplo, el basado en redes neuronales) su funcionamiento es indescifrable por naturaleza y ni siquiera los expertos que han diseñado el sistema pueden identificar los pasos a través de los cuales ha llegado a una decisión porque los modelos predictivos o las reglas de decisión emergen y cambian automáticamente a partir del análisis masivo de datos que analiza. A esto se le conoce como el «efecto caja negra» (*black box*), es decir, la imposibilidad de obtener una explicación en términos causales acerca de cómo el algoritmo ha generado cierta información de salida o qué combinación de *inputs* han contribuido a ella.

En este tipo de sistemas, ni siquiera el acceso al código fuente permitirá desvelar los pasos seguidos por los algoritmos para realizar sus predicciones. Por ello, algunos autores hablan de la «falacia de la transparencia», ya que no siempre que se consigue acceder a la caja negra queda garantizada la comprensibilidad de los resultados, ni salvaguardados adecuadamente los derechos de los interesados¹⁴⁸.

A esta opacidad estructural o técnica de los sistemas se une también la opacidad jurídica, en la medida en que el acceso al algoritmo podrá denegarse al interesado si se considera protegido por derechos de propiedad intelectual o como secreto empresarial. Por tanto, aunque se utilicen algoritmos de «caja blanca» o explicables, el interesado podría encontrar muchos problemas para obtener una motivación de la decisión que le afecta. A lo que habría que sumar otro escollo a la transparencia que a su vez dificultaría la averiguación del origen de la discriminación algorítmica, que sería la propiedad de los datos utilizados para entrenar al algoritmo¹⁴⁹. Esta cuestión no es baladí por cuanto lo realmente útil para detectar el sesgo podría ser el acceso a los datos que han servido de soporte al aprendizaje

¹⁴⁸ SOLAR CAYÓN, J.I., *op.cit.*, p. 148. Explica este autor que en estos casos las predicciones son «el resultado de la compleja interacción de una red de millones de neuronas artificiales distribuidas en múltiples capas o niveles de análisis que ponderan simultáneamente el valor predictivo de cientos o miles de factores potenciales y de sus múltiples posibilidades combinatorias, lo que imposibilita identificar el peso de cada factor en los resultados». Por ello, son indescifrables por naturaleza ya que sus decisiones no son el resultado de ejecutar una serie de fórmulas o reglas de decisión previamente programadas, sino de identificar patrones ocultos imperceptibles para los humanos (y en ocasiones sin sentido) y correlaciones entre los datos masivos que analizan, a partir de los cuales generan automáticamente los modelos predictivos o reglas de decisión a aplicar a nuevos conjuntos de datos que además pueden ir cambiando automáticamente.

¹⁴⁹ Destaca este problema, por ejemplo, RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», *cit.*, pp. 448 y ss. Sobre el derecho de acceso a los datos de entrenamiento de sistema, puede verse ampliamente, LÓPEZ-TARRUELLA MARTÍNEZ, A., *Propiedad intelectual e innovación basada en los datos*, Navarra, 2021, pp. 192 y ss.

del algoritmo. Esto ocurriría en los casos de algoritmos predictivos cuyo aprendizaje se basa precisamente en los datos aportados, a diferencia de lo que ocurriría en los algoritmos no predictivos en los que los datos de alimentación serían secundarios; en estos últimos, sí que podría ser útil acceder al código fuente o a la secuencia de programación para averiguar el origen de la decisión, detectando así fallos o errores en alguno de sus pasos¹⁵⁰.

Esta doble opacidad, que dificulta tanto el *acceso* como la *comprensión* del sistema, pone en tela juicio el derecho a obtener una resolución o decisión motivada y la posibilidad de recurrirla, generando indefensión para el afectado y un grave riesgo de arbitrariedad, sobre todo cuando estos sistemas son utilizados por la Administración puesto que en el ámbito público las exigencias de transparencia en los procesos decisorios se acrecientan.

Hasta hace poco tiempo, los desarrolladores de estos sistemas se enfocaban en maximizar la capacidad predictiva de los algoritmos, incluso si esto implicaba una mayor complejidad y opacidad del sistema. Sin embargo, recientemente ha surgido una tendencia hacia la creación de modelos más simples, transparentes y explicables, que cumplan con los requisitos legales y aumenten la confianza de la ciudadanía. De hecho, contamos a día de hoy con toda un área de trabajo denominada *IA Explicable* (*Explainable AI –XAI–*) dirigida a encontrar métodos para que los expertos humanos puedan comprender por qué el sistema se comportó de una determinada manera. Es decir, se trata de encontrar un adecuado equilibrio entre precisión y explicabilidad para conseguir que la tecnología resulte fiable.

¹⁵⁰ RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», *cit.*, p. 439.

Capítulo II

La discriminación algorítmica a grupos vulnerables

1. VENTAJAS Y RIESGOS DEL USO DE SISTEMAS DE IA PARA LA TOMA DE DECISIONES

Son innegables e innumerables las ventajas que proporcionan los sistemas inteligentes para la realización de múltiples tareas humanas como la toma de decisiones. Esto se debe a su creciente capacidad computacional, es decir, a su posibilidad de almacenar y tratar ingentes cantidades de datos inasumibles para el cerebro humano y a una velocidad sin precedentes, a la cada vez mayor existencia y disponibilidad de datos y a la sofisticación de los algoritmos de aprendizaje automático¹⁵¹. Todo ello ha permitido un exponencial avance de la IA y que su uso se haya generalizado en prácticamente todos los sectores de la sociedad buscando velocidad, precisión y objetividad en la toma de decisiones.

De hecho, para un sector doctrinal, el hecho de que algunos de estos sistemas alcancen un alto grado de precisión o fiabilidad puede contribuir a reducir daños en muchos sectores, como la automoción o la salud. Sobre esto se viene manteniendo que, puesto que la mayoría de accidentes de circulación

¹⁵¹ La Comisión Europea, en su Comunicación al Parlamento Europeo, al Consejo Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones titulada «Inteligencia artificial para Europa», ya destacaba esta idea al señalar que «se han conseguido avances significativos en estas tecnologías gracias al *uso de grandes conjuntos de datos y de una capacidad informática sin parangón*». También es unánime entre los autores considerar que estos elementos han sido determinantes para el progreso de la IA. En esta línea, entre otros, RANDO RAMIREZ https://repositori.upf.edu/bitstream/handle/10230/47169/Rando_Sant-Jordi_2020.pdf?sequence=1&isAllowed=y, para quien la increíble potencia de cálculo que poseen los ordenadores y las enormes bases de datos disponibles nos permiten crear poderosas herramientas que nos ayudan a tomar decisiones más justas; HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, p. 39; HERRERA DE LAS HERAS, R., «Protección de datos e inteligencia artificial», *cit.*, p. 644.

son debidos a errores humanos, la implementación de los coches autónomos contribuirá a reducirlos, y por ende, disminuirán los daños en este ámbito; de forma similar, en el ámbito de la salud, si la IA permite contar con herramientas más fiables o eficaces para el diagnóstico o tratamiento de enfermedades, es lógico pensar que se podrían llegar a salvar muchas más vidas que sin ellas¹⁵². Un ejemplo de esta mayor fiabilidad o precisión de los sistemas de IA en el ámbito médico podría ser el de un sistema IA de *deep learning* desarrollado por Google, que logró superar a los radiólogos diagnosticando cáncer de pulmón en las primeras etapas de la enfermedad gracias al análisis previo de más de 42.000 tomografías¹⁵³.

No es de extrañar, por todo ello, que las empresas recurran a estos sistemas para mejorar su productividad, optimizar recursos y ser más competitivos, ya que es indiscutible que la automatización en la toma de decisiones (incluida la elaboración de perfiles) puede suponer un ahorro importante de tiempo y esfuerzo para los trabajadores y de dinero para el empresario, ya que la máquina realizará la tarea, no sólo de forma más rápida que el humano, sino también –en teoría– de forma más eficiente, neutral y objetiva, al ser capaz de procesar infinidad de datos de forma casi instantánea, de encontrar correlaciones entre dichos datos y de extraer conclusiones carentes de subjetividad debido a que la máquina «ni siente, ni padece» y no se verá influenciada en sus decisiones, ni en su nivel de productividad, por el estado físico o anímico en que se encuentre ese día¹⁵⁴.

¹⁵² En este sentido, RUBÍ PUIG, A., «Retos de la inteligencia artificial y adaptabilidad del derecho de daños», *Retos jurídicos de la inteligencia artificial*, coord. por Agustí Cerrillo i Martínez, Miquel Peguera Poch, 2020, p. 60; ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...*cit., p. 25; RACHUM-TWAI, O., «Whose robot is it anyway?: Liability for artificial-intelligence-based robots», 2020 *Univ. Ill. L.Rev.* 4, p. 1158. Recuperado de <https://www.illinoislawreview.org/wp-content/uploads/2020/08/Rachum-Twaig.pdf>.

¹⁵³ Sobre este sistema pone de manifiesto GARCÍA TERUEL, R.M., *op.cit.*, p. 1012, que tenía un 11% menos de falsos positivos y un 5% menos de falsos negativos que cuando el diagnóstico se llevaba a cabo por profesionales de la medicina. También puede verse sobre este sistema de *deep learning*: <https://www.xataka.com/inteligencia-artificial/ia-google-logra-superar-a-radiologos-diagnosticando-cancer-pulmon-primeras-etapas-enfermedad>.

¹⁵⁴ Señala ALAMEDA CASTILLO, M.T., *op.cit.*, p. 13, que todavía habría más, ya que, por ejemplo, a la hora de hacer una selección de personal a través de un sistema IA el algoritmo no sólo sería capaz de buscar las características deseadas en los currículums de los aspirantes al puesto y de seleccionar los candidatos más adecuados, sino que, a partir de su propio aprendizaje, determinaría las características idóneas para ese puesto, por repetición y por asociación. También mantiene RIVAS VALLEJO, P. «Sesgos de automatización...», *cit.*, p. 52, que los modelos de decisión automatizados podrían presentarse como alternativa más objetiva, rápida y eficaz a los procesos tradicionales de toma de decisiones en la medida en que su falta de interés e imparcialidad mejoraría la objetividad de la respuesta. Además, serían más neutrales, primando la eficacia de estos modelos por la mejora que produce en los tiempos de respuesta y en la minimización de sesgos.

Por tanto, con estos sistemas inteligentes teóricamente se reduce el riesgo de cometer errores en las decisiones adoptadas y se consiguen decisiones más justas en la medida en que se evita que se den dos respuestas diferentes ante una misma situación. Es decir, se consigue una consistencia y uniformidad en las decisiones que se ha destacado desde el punto de vista jurídico como una de las bondades más significativas de dichos sistemas en relación con las decisiones humanas, las cuales son subjetivas y están condicionadas por sesgos personales y diferencias en la interpretación de una misma información¹⁵⁵.

Esta supuesta imparcialidad y objetividad de los sistemas de IA para la toma de decisiones ha sido puesta de manifiesto en la *Estrategia de Inteligencia Artificial de la Comunidad Valenciana* donde se destaca que «los beneficios de aplicar la IA se reflejarán en una *mejor toma de decisiones* al tener en cuenta todas las posibles variables y contrastar un mayor número de datos. *Al estar basadas en algoritmos que no se ven afectados por subjetividades personales, las decisiones son más objetivas*. También son más rápidas porque la capacidad de cómputo supera a la capacidad de la inteligencia humana»¹⁵⁶.

Siendo esto así, podría llegar a pensarse (como de hecho ocurre) que el uso de estas tecnologías tan capaces de analizar la realidad social y predecir en base a ella y con tanta precisión hechos o comportamientos humanos futuros en base a datos pasados, con una rapidez, eficacia, neutralidad y objetividad inalcanzables para los humanos, no tendría más que beneficios y nulas posibilidades de dañar a las personas. Sin embargo, nada más lejos de la realidad. Al margen de los casos más flagrantes de daños a la vida, salud o seguridad de las personas (piénsese, por ejemplo, en los accidentes ocasionados por los vehículos autónomos¹⁵⁷, o la IA de Facebook que tuvo que ser desconectada porque inventó su propio idioma indecifrible para los humanos e ignoraba las órdenes de sus programadores¹⁵⁸), los sistemas inteligentes implican ries-

¹⁵⁵ Por todos, SOLAR CAYÓN, J.I., *op.cit.*, p. 136 y pp. 153-154. Pone de relieve este autor que aunque se ha mantenido que estos sistemas son más fiables porque permiten llegar a decisiones más objetivas y uniformes que las decisiones del juzgador humano, las cuales pueden ser subjetivas, discrecionales y condicionadas por sesgos, no hay evidencia científica firme para sostener que las predicciones realizadas por estos sistemas son más correctas que las realizadas por los jueces humanos.

¹⁵⁶ GVA, 2018, p. 5. Las cursivas son nuestras.

¹⁵⁷ De sobra conocidos son los casos de atropellos por parte de coches autónomos, como sucedió en el primer accidente mortal del vehículo autónomo de Uber en Arizona en 2018, al que le han sucedido otros muchos. También se han dado casos de coches Tesla que presentan fallos del piloto automático para detectar señales, o que se bloquen y arden con el conductor dentro. Pueden verse, entre otras muchas, las siguientes páginas que informan sobre estos sucesos: <https://www.elmundo.es/motor/2022/06/19/62aeba79fd4c408b4588.html>; <https://motor.elpais.com/videos/2/ultimo-fallo-de-un-tesla-se-bloquea-y-arde-con-el-conductor-dentro/>

¹⁵⁸ <https://www.fayerwayer.com/internet/2023/01/10/el-adios-de-alice-y-bob-que-ocurrio-con-la-inteligencia-artificial-que-creo-meta-y-desarrollo-su-propio-idioma/>

gos nada desdeñables para otros derechos fundamentales como la intimidad, la protección de datos personales o la igualdad y no discriminación, que a diferencia de los primeros pasan mucho más desapercibidos para los ciudadanos. Esto obliga a instituciones e investigadores a advertir sobre ellos y a poner todos los medios disponibles para detectarlos y evitarlos, o al menos minimizarlos, con el fin de reducir también potenciales daños¹⁵⁹.

Dichos riesgos, que están directamente vinculados con las características intrínsecas de estos sistemas¹⁶⁰, han sido reconocidos en prácticamente todas las resoluciones, propuestas e iniciativas que se han ido adoptando en materia de IA en los últimos años, tanto en el ámbito internacional como nacional, y sobre todo, europeo. Ya en el año 2017, el Parlamento Europeo alertaba de ellos en varias de sus resoluciones. Por ejemplo, en su *Resolución de 16 de febrero de 2017 sobre normas de derecho civil sobre robótica* con recomendaciones a la Comisión¹⁶¹, reconocía que el aprendizaje automático ofrecía enormes ventajas económicas a la sociedad al mejorar enormemente la capacidad de analizar datos, pero también planteaba retos a la hora de velar por la no discriminación, las garantías procesales, la transparencia y la inteligibilidad de los procesos decisorios¹⁶². Y en su *Resolución de 14 de marzo de 2017, sobre las implica-*

¹⁵⁹ De ellos advertía ya la Comisión Europea en su Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones (COM/2019/168 final), donde afirmaba que «La IA trae retos, ya que permite a las máquinas “aprender” y tomar decisiones y ejecutarlas sin intervención humana. Ahora bien, las decisiones adoptadas por algoritmos pueden dar datos incompletos y, por tanto, no fiables, que puede ser manipulados, sesgados o simplemente estar equivocados». También en el *Libro Blanco sobre IA*, cit., p. 13, se pone de manifiesto esta realidad cuando se afirma que, aunque la IA puede ofrecer muchas ventajas, por ejemplo, mejorando la seguridad de los productos y los procedimientos, también puede resultar nociva y puede conllevar daños tanto materiales (para la seguridad y la salud de las personas, con consecuencias como la muerte, y menoscabos al patrimonio) como inmateriales (pérdida de privacidad, limitaciones del derecho de libertad de expresión, dignidad humana, discriminación en el acceso al empleo, etc.) y pueden estar vinculados a una gran variedad de riesgos. El marco regulador debe centrarse en cómo minimizar los distintos riesgos de sufrir daños, especialmente los más significativos. Los principales riesgos relacionados con el uso de la inteligencia artificial afectan a la aplicación de las normas diseñadas para proteger los derechos fundamentales (como la protección de los datos personales y la privacidad, o la no discriminación) y la seguridad, así como a las cuestiones relativas a la responsabilidad civil.

¹⁶⁰ Ya expuestas en los apartados anteriores. En las conclusiones de la *Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno*, el Consejo de Europa considera que para garantizar la compatibilidad de la IA con los derechos fundamentales y facilitar la aplicación de las normas jurídicas, es preciso afrontar retos como la *opacidad, la complejidad, el sesgo, cierto grado de imprevisibilidad y un comportamiento parcialmente autónomo* de estos sistemas.

¹⁶¹ Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103(INL)).

¹⁶² Vid., Considerando letra h). También en el apartado sobre Principios éticos señala que «el potencial de empoderamiento que encierra el recurso a la robótica se ve matizado por una serie de

ciones de los macrodatos en los derechos fundamentales: privacidad, protección de datos, no discriminación, seguridad y aplicación de la ley, insistía en la misma idea¹⁶³.

También la Comisión Europea da buena cuenta de todo esto en su *Libro Blanco sobre IA*, al definir los problemas que plantean estos sistemas y señala que «[e]l uso de la inteligencia artificial puede afectar a los valores sobre los que se fundamenta la UE y provocar la conculcación de derechos fundamentales, como la *libertad de expresión, la libertad de reunión, la dignidad humana, la ausencia de discriminación por razón de sexo, raza u origen étnico, religión o credo, discapacidad, edad u orientación sexual*, y, en su aplicación en determinados ámbitos, la *protección de los datos personales y de la vida privada, el derecho a una tutela judicial efectiva y a un juicio justo, o la protección de los consumidores*».

Y el recientemente aprobado RIA, que también va dirigido a conseguir una IA segura, fiable y respetuosa con los derechos fundamentales y los principios éticos, desde su versión inicial contiene en sus Considerandos una mención expresa a los derechos fundamentales que pueden verse amenazados por los distintos usos de esta tecnología, entre ellos, la *dignidad humana, la protección de datos personales e intimidad, la igualdad y no discriminación entre mujeres y hombres*, prestando especial atención a los mecanismos para evitar el menoscabo al derecho a la *libertad de expresión y reunión*. Asimismo, hace especial hincapié en los *derechos relacionados con la justicia*, al ser el vehículo para poder garantizar el resto de los derechos fundamentales recogidos en la Carta. De ahí la especial alusión al derecho a la *tutela judicial efectiva* y al *juez imparcial*, así como a la *presunción de inocencia*¹⁶⁴. Para conseguir dicha seguridad y confianza en la tecnología, incorpora un marco jurídico con normas armonizadas en materia de IA orientadas a impulsar su desarrollo, utilización y adopción en el mercado interior y a ofrecer, al mismo tiempo, un nivel elevado de protección de los intereses públicos, como la salud y la seguridad, y los derechos fundamentales reconocidos y protegidos por el Derecho de la Unión. Dicha iniciativa pretendía, desde sus orígenes, ser coherente con la Carta de los Derechos Fundamentales de la Unión Europea y el Derecho derivado de la Unión vi-

tensiones o posibles riesgos y que debe ser evaluado detenidamente a la luz de la seguridad y la salud humanas; la libertad, la intimidad, la integridad y la dignidad; la autodeterminación y la no discriminación, y la protección de los datos personales» (nº 10).

¹⁶³ Vid., Considerando I: «Considerando que los macrodatos presentan el potencial mencionado anteriormente para los ciudadanos, el mundo académico y la comunidad científica y los sectores público y privado, pero que también conllevan riesgos significativos, concretamente en lo que se refiere a la protección de derechos fundamentales como el derecho a la privacidad, la protección y la seguridad de los datos, además de la libertad de expresión y la no discriminación, garantizados por la Carta de los Derechos Fundamentales y la legislación de la Unión».

¹⁶⁴ Vid., el apartado 3.5 de la versión original del RIA, de abril de 2021 y su Considerando 28.

gente en materia de protección de datos, protección de los consumidores, no discriminación e igualdad.

Por su parte, el Consejo de Europa realizó un exhaustivo estudio titulado *Algoritmos y Derechos humanos* donde puso de manifiesto que el elenco de derechos fundamentales que podían verse afectados por estos sistemas de IA era muy extenso¹⁶⁵. Y asimismo la Agencia de los Derechos Fundamentales (FRA) manifestó en su *Informe sobre IA y derechos fundamentales*, que a la hora de utilizar la IA era imprescindible considerar una gran variedad de derechos que podían variar en función de la tecnología empleada y su ámbito de utilización. Así, «además de los derechos relativos a la *privacidad, la protección de datos, la igualdad y la no discriminación, y el acceso a la justicia*, podrían considerarse otros derechos, como, por ejemplo, la *dignidad humana, el derecho a la seguridad social y la asistencia social, el derecho a una buena administración* (especialmente relevante para el sector público) y la *protección de los consumidores* (particularmente importante para las empresas)»¹⁶⁶.

1.1. LA DISCRIMINACIÓN ALGORÍTMICA COMO UNO DE LOS PRINCIPALES RIESGOS DE LAS DECISIONES DE LOS SISTEMAS DE IA

Por lo que se refiere en concreto al derecho a la igualdad de trato y la no discriminación, en numerosos documentos nacionales, europeos e internacionales ha sido constante el reconocimiento de que este derecho fundamental, junto con el de protección de datos personales, es especialmente vulnerable a través de esta tecnología y merecedor de una tutela especial, aumentado el riesgo de que ambos derechos sean vulnerados de forma proporcional al grado de autonomía del sistema utilizado, ya que el humano puede llegar a perder totalmente la dirección y control sobre las decisiones adoptadas, sin que además se pueda llegar a saber en base a qué motivos el sistema decidió como lo hizo¹⁶⁷.

¹⁶⁵ En <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5>. También pueden verse la Recomendación CM/Rec(2020)1 de 8 de abril de 2020, del Comité de Ministros del Consejo de Europa, sobre *impacto en los derechos humanos de los sistemas algorítmicos*, disponible en https://search.coe.int/cm/pages/result_details.aspx?objectId=09000016809e1154, y la Recomendación CM/Rec(2021)8, sobre *protección de las personas en relación con el tratamiento automatizado de datos en el contexto de la elaboración de perfiles*, disponible en https://search.coe.int/cm/pages/result_details.aspx?ObjectId=0900001680a46147.

¹⁶⁶ Recuperado de https://fra.europa.eu/sites/default/files/fra_uploads/fra-2021-artificial-intelligence-summary_es.pdf, p.5.

¹⁶⁷ Por todos, puede verse la especial referencia que se hace a estos dos derechos fundamentales en el *Libro Blanco sobre IA*, cit., p. 13. También la doctrina suele destacar entre los riesgos más importantes de la IA, el sesgo o la discriminación algorítmica. En este sentido puede verse, por todos, el trabajo de GINÈS i FABRELLAS, A., «Sesgos discriminatorios en la automatización de decisiones en

De hecho, en numerosos instrumentos se aborda explícita y concretamente el sesgo algorítmico como uno de los riesgos más relevantes de la IA que hay que combatir. En sus *Directrices éticas para una IA fiable*, el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial identificó siete requisitos clave que debían cumplir los sistemas IA para ser considerados fiables, siendo el quinto de ellos la *necesidad de evitar el sesgo algorítmico injusto*. Tal como señala, «en el contexto de la IA, la igualdad implica que el funcionamiento de este tipo de sistemas no debe generar resultados injustamente sesgados (por ejemplo, los datos utilizados para la formación de los sistemas de IA deben ser lo más inclusivos posibles, de forma que estén representados los diferentes grupos de población). Esto también requiere un adecuado respeto de las personas y grupos potencialmente vulnerables, como los trabajadores, las mujeres, las personas con discapacidad, las minorías étnicas, los niños, los consumidores u otras personas en riesgo de exclusión»¹⁶⁸.

Además, la ausencia de sesgos y discriminación se configura como un *principio ético* de obligado cumplimiento para las tecnologías de alto riesgo en la *Resolución del Parlamento Europeo, de 20 de octubre de 2020, con recomendaciones destinadas a la Comisión sobre un marco de los aspectos éticos de la inteligencia artificial, la robótica y las tecnologías conexas (2020/2012(INL))* (en adelante, Resolución sobre aspectos éticos de la IA 2020) que concibe la discriminación como una de las posibles causas en la producción de «lesión o daño» en el art. 4 del Reglamento que la acompaña¹⁶⁹, estando presente en toda la propuesta el de-

el ámbito laboral: evidencias de la práctica», en RIVAS VALLEJO, P. (dir), *Discriminación algorítmica en el ámbito laboral*, 2022; o los trabajos de SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación: aproximación y propuestas generales», *Revista General de Derecho Administrativo*, N.º. 56, 2021, «Decisiones automatizadas: problemas y soluciones jurídicas. Más allá de la protección de datos», *Revista de Derecho Público: teoría y método*, N.º13, 2021, pp. 85-127, y «La aplicación del marco jurídico europeo en materia de igualdad y no discriminación al uso de aplicaciones de Inteligencia Artificial», en *Nuevas normatividades: inteligencia artificial, derecho y género* (coord. por P.R. Bonorino Ramírez / R. Fernández Acevedo / P. Valcárcel Fernández), Navarra, 2022, pp. 63-88.

¹⁶⁸ P. 13.

¹⁶⁹ El art. 4 letra n) define la «lesión o daño» como «toda lesión física o mental y todo daño material o inmaterial, incluso cuando hayan sido causados por incitación al odio, sesgo, discriminación o estigmatización, tales como pérdidas financieras o económicas, pérdida de empleo o de oportunidades educativas, restricción indebida de la libertad de elección o de expresión, pérdida de la privacidad y cualquier infracción del Derecho de la Unión que sea perjudicial para una persona». También en la *Resolución del Parlamento Europeo, de 20 de octubre de 2020, con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial (2020/2014(INL))* (en adelante, Resolución sobre RC de 2020), se hace referencia como posible daño (grave) ocasionado por los sistemas de IA a «la toma de decisiones sesgadas en materia de seguro de enfermedad, concesión de crédito, autos judiciales, contratación o empleo» (Considerando 3). También en la *Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno* del Consejo de Europa recoge entre los cinco principios sustanciales y metodológicos aplicables al procesamiento

recho a la no discriminación como uno de los pilares básicos del Derecho de la Unión¹⁷⁰.

Aunque dicha Propuesta de Reglamento no salió adelante, se avanzaban en ella cuestiones importantes en relación con la discriminación algorítmica. En este sentido, se recogen algunas de las *causas o factores* que pueden provocar el sesgo algorítmico (Considerando 22), o cuándo debe considerarse sesgado o discriminatorio un programa informático, algoritmo o dato utilizado o producido por la inteligencia artificial, la robótica y las tecnologías conexas, poniendo como ejemplos los casos en que no arrojen resultados óptimos en relación con cualquier persona o grupo de personas, sobre la base de una percepción personal o social prejuiciosa y del posterior tratamiento de los datos relativos a sus características (Considerando 23), o cuando produzcan resultados con unos efectos negativos desproporcionados que den lugar a un tratamiento diferente de una persona o grupo de personas, que las ponga incluso en una situación de desventaja en comparación con otras personas, sobre la base de motivos tales como sus características personales, sin una justificación objetiva o razonable e independientemente de cualquier reivindicación de neutralidad de las tecnologías en cuestión (Considerando 24).

Se destaca también en ese texto que cuando a través de estas tecnologías se lleve a cabo un tratamiento de datos personales, incluidos los datos personales derivados de datos no personales y de datos biométricos, deberá respetarse el Reglamento (UE) 2016/679 y la Directiva 2002/58/CE. Sobre esto, la AEPD considera la discriminación derivada de los sesgos introducidos en los procesos algorítmicos como el principal riesgo derivado del tratamiento de datos personales con los que funcionan los sistemas de IA¹⁷¹. Y también el GT29 subraya el peligro de que tanto la elaboración de perfiles como las decisiones automatizadas perpetúen los estereotipos existentes y la segregación social y lleven a discriminación injustificada¹⁷².

Reconoce también la no discriminación como uno de los principios fundamentales que deben guiar el desarrollo de la IA, la UNESCO, en su *Recomendación sobre la Ética de la Inteligencia Artificial*, de noviembre de 2021,

automatizado de decisiones y datos judiciales, basados en técnicas de inteligencia artificial, el «principio de no discriminación» con el que se trata de prevenir específicamente el desarrollo o intensificación de cualquier discriminación entre individuos o grupos de individuos.

¹⁷⁰ Sobre esta propuesta de Reglamento sobre los aspectos éticos de la IA, puede verse LAZCOZ MORATINOS, G., «Análisis de la propuesta de Reglamento sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas», *IUS ET SCIENTIA: Revista electrónica de Derecho y Ciencia*, Vol. 6, N.º. 2, 2020, pp. 26-41.

¹⁷¹ Informe sobre la *Adecuación al RGPD de tratamientos que incorporan IA*, cit., p. 30.

¹⁷² Vid., *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679*.

primer acuerdo mundial sobre un marco ético que define los valores y principios comunes que deben guiar y garantizar un desarrollo saludable de esta tecnología. Se afirma expresamente que es necesario evitar los sesgos que pueden exacerbar discriminaciones existentes, desigualdad, brechas digitales y exclusión y suponer una amenaza para la diversidad cultural, social y biológica, así como generar divisiones sociales o económicas.

Particular atención se presta al riesgo de discriminación algorítmica en el RIA, cuya versión inicial declaraba expresamente que uno de los objetivos del establecimiento de requisitos específicos para los sistemas de alto riesgo era «*reducir al mínimo el riesgo de discriminación algorítmica, en particular en lo tocante al diseño y la calidad de los conjuntos de datos empleados para desarrollar sistemas de IA*», reconociendo con ello las principales causas que pueden provocarla¹⁷³.

De hecho, el peligro de que los sistemas de IA contengan sesgos y generen discriminación injusta es tomado en consideración a lo largo de toda la Propuesta para clasificarlos dentro de uno u otro nivel de riesgo. Así, por ejemplo, para prohibir los sistemas de puntuación o clasificación ciudadana, o los de identificación biométrica remota «en tiempo real» de personas físicas en espacios de acceso público con fines de garantía del cumplimiento del Derecho (con excepciones), respecto de los que se destaca que las imprecisiones técnicas que conllevan pueden provocar resultados sesgados y tener consecuencias discriminatorias, lo que cobra especial importancia en lo que respecta a la edad, la etnia, el sexo o la discapacidad¹⁷⁴. También respecto a los sistemas de IA que se utilizan en la educación o la formación profesional, y en especial aquellos que determinan el acceso o distribuyen a las personas entre distintas instituciones educativas y de formación profesional o aquellos que evalúan a las personas a partir de pruebas realizadas en el marco de su educación o como condición necesaria para acceder a ella, ya que pueden decidir la trayectoria formativa y profesional de una persona y, en consecuencia, afectar a su capacidad para asegurar su subsistencia. Respecto de ellos se advierte que su mal diseño o uso incorrecto puede vulnerar el derecho a la educación y la formación, y el derecho a no sufrir discriminación, además de perpetuar patrones históricos de discriminación y se califican como de alto riesgo¹⁷⁵. Similarmente respecto a los sistemas de IA que se emplean en el ámbito laboral, sobre todo para la contratación y la selección de personal, se reconoce que los mismos pueden perpetuar patrones históricos de discriminación, por ejemplo, contra las mujeres, ciertos grupos de edad, personas con discapacidad o personas de orígenes raciales o étnicos concretos o con una

¹⁷³ Ap.1.2.

¹⁷⁴ Vid., Considerandos 32.

¹⁷⁵ Vid., Considerando 56.

orientación sexual determinada, y por ello se califican como de alto riesgo¹⁷⁶. Y lo mismo ocurre con aquéllos sistemas de IA usados para evaluar la calificación crediticia o solvencia de personas físicas (que deciden si las personas pueden acceder a recursos financieros o servicios esenciales como la vivienda, la electricidad y los servicios de telecomunicaciones), los cuales pueden discriminar a personas o grupos y perpetuar patrones históricos de discriminación, por ejemplo, por motivos de origen racial o étnico, discapacidad, edad u orientación sexual, o generar nuevas formas de discriminación¹⁷⁷. Puede inferirse con facilidad que, para el RIA, la discriminación es especialmente crítica en todos estos sectores donde las decisiones algorítmicas pueden afectar significativamente las oportunidades de las personas.

La preocupación por el riesgo de que los sistemas de IA discriminen injustamente, se observa también en el reciente *Convenio Marco del Consejo de Europa sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho*, firmado en 2024, que es el primer acuerdo internacional jurídicamente vinculante en materia de IA¹⁷⁸. Se reconoce en su Preámbulo que «los riesgos de discriminación en contextos digitales, en particular los que implican sistemas de inteligencia artificial, y su posible efecto de crear o agravar desigualdades, incluidas las que sufren las mujeres y las personas en situaciones vulnerables, en relación con el disfrute de sus derechos humanos y su participación plena, igualitaria y efectiva en los asuntos económicos, sociales, culturales y políticos», es uno de los retos específicos que se pretende abordar.

La relevancia del problema de los sesgos o discriminación algorítmica en la toma automatizada de decisiones es de tal magnitud que ha sido la primera cuestión relacionada con la IA objeto de regulación positiva en nuestro país. En concreto, en la *Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación*¹⁷⁹, cuyo art. 3 señala expresamente entre los ámbitos de aplicación de la norma el de la «Inteligencia Artificial y gestión masiva de datos, así como

¹⁷⁶ Vid., Considerando 57.

¹⁷⁷ Vid., Considerando 58. No obstante, como se puntualiza en este Considerando, no deben considerarse de alto riesgo en virtud del presente Reglamento «los sistemas de IA previstos por el Derecho de la Unión con vistas a detectar fraudes en la oferta de servicios financieros y, a efectos prudenciales, para calcular los requisitos de capital de las entidades de crédito y las empresas de seguros».

¹⁷⁸ Este Convenio fue firmado por la Comisión Europea el 5 de septiembre de 2024 en nombre de la UE. Para mayor información sobre el mismo, nos remitimos al apartado 2.2 del Capítulo IV de este trabajo.

¹⁷⁹ Sin perjuicio de lo afirmado, algún autor considera que la primera regulación de la IA en nuestro país la llevó a cabo la *Ley 1/2022, de 13 de abril, de Transparencia y Buen Gobierno de la Comunitat Valenciana*. Así, COTINO HUESO, L., «Cuándo deben crearse registros y dar transparencia a los algoritmos y sistemas de inteligencia artificial públicos», en COTINO, L./ CASTELLANOS, J., *Algoritmos abiertos y que no discriminen en el sector público*, Valencia, 2023, pp. 137-172. Sin embargo, el problema de los sesgos que a nosotros interesa no se contempla en dicha Ley de transparencia.

otras esferas de análoga significación», estableciendo en su art. 23 que las administraciones públicas favorecerán la puesta en marcha de mecanismos para que los algoritmos involucrados en la toma de decisiones tengan en cuenta criterios de *minimización de sesgos, transparencia y rendición de cuentas*; que se promoverá la realización de *evaluaciones de impacto* que determinen el posible sesgo discriminatorio; que se priorizarán la *transparencia en el diseño* y la implementación y la *capacidad de interpretación de las decisiones* adoptadas por los mismos; y que, tanto las administraciones públicas como las empresas privadas, *promoverán el uso de una IA ética, confiable y respetuosa con los derechos fundamentales*, siguiendo especialmente las recomendaciones de la Unión Europea en este sentido, promoviendo también un *sello de calidad* de los algoritmos. Por tanto, se asumen las exigencias ya recogidas por el Grupo de Expertos en sus *Directrices éticas para una IA fiable* de 2019, y se fijan como criterios rectores en el uso de la IA, la minimización de sesgos, transparencia y rendición de cuentas, transparencia en diseño y la implementación, interpretabilidad, así como la promoción de la ética y confiabilidad.

Además, en base a esta Ley, algunas Comunidades Autónomas han adoptado normas para impulsar la IA en sus territorios. Es el caso de Extremadura con el Decreto-ley 2/2023, de 8 de marzo, de medidas urgentes de impulso a la inteligencia artificial. También en Galicia se aprobó, el 2 de diciembre de 2024, un Proyecto de Ley para el desarrollo e impulso de la IA en dicha Comunidad cuyo objetivo es garantizar un desarrollo y uso ético, seguro, fiable y centrado en las personas. Para ello, cuando la Administración utilice esta tecnología deberá contar con un informe de proyecto de IA que, además de describir su funcionamiento y características detalladas, incluirá una «evaluación de impacto en los derechos fundamentales» de acuerdo con el modelo de la Oficina Europea de IA. También se prevé la «supervisión humana» en todos los procedimientos administrativos que usen esta tecnología, siempre que sea necesaria o solicitada por cualquier persona implicada en el trámite¹⁸⁰.

Por otro lado, entre los instrumentos de *soft law*, debe mencionarse la *Carta de Derechos Digitales* adoptada por el Gobierno español el 14 de julio de 2021, donde se contempla la prohibición de discriminación algorítmica sea cual sea su origen o causa¹⁸¹. Y a nivel europeo también contamos con una *Declaración sobre los Derechos y Principios Digitales para la Década Digital* que proclama el derecho a la igualdad y no discriminación, la inclusión y la accesibilidad universal en el entorno digital¹⁸². Además, respecto a las interacciones

¹⁸⁰ <https://www.xunta.gal/es/notas-de-prensa/-/nova/007850/rueda-anuncia-que-galicia-convierte-primer-region-europea-que-regula-por-ley>.

¹⁸¹ Vid., apartado XVIII.6.

¹⁸² Vid., Capítulo III de la *Declaración Europea sobre los Derechos y Principios Digitales para la Década Digital* (2023/C 23/01). Y en el Capítulo II se establece que las soluciones tecnológicas deberán respe-

con algoritmos y sistemas de IA declara el compromiso de *velar por la transparencia*, manteniendo que los sistemas algorítmicos *no deben conllevar discriminación ilegal* y que hay que permitir la *supervisión humana* en los resultados que afecten a personas.

Todo lo anterior refleja que los distintos organismos y autoridades nacionales e internacionales han tomado conciencia de los riesgos que conlleva la IA con el fin de adoptar las medidas preventivas o reparadoras oportunas. No obstante, a pesar del gran avance que ello supone para la seguridad y salvaguarda de los derechos fundamentales, no es suficiente. También la ciudadanía debe ser plenamente consciente de esta situación. La responsabilidad de eliminar o mitigar sesgos en los algoritmos recae en particulares, empresas y gobiernos. Es importante, por ello, que desde todos los ámbitos se haga hincapié tanto en las fortalezas como en las debilidades de estas tecnologías. En este sentido, al menos desde el ámbito jurídico, vamos en la buena dirección en tanto en cuanto la práctica totalidad de autores que se ha sumergido en temas relacionados con la IA, inciden en la dicotomía de *riesgos-beneficios* que la misma encierra, sin perjuicio de que para un sector doctrinal sea preciso poner el acento en las ventajas y de que, para otros, los riesgos sean tan relevantes que pueden poner en entredicho sus ventajas¹⁸³.

Por nuestra parte, no albergamos dudas de que los beneficios que puede proporcionar la IA superan sobradamente sus inconvenientes¹⁸⁴. Baste citar

tar los derechos de las personas, permitiendo su ejercicio y promoviendo la inclusión, prestándose especial atención a las personas mayores, con discapacidad o marginadas, vulneradas o privadas de derechos (incluyendo a quienes actúen en su nombre).

¹⁸³ Como explica NÚÑEZ ZORRILLA, M.C., «Los nuevos retos de la Unión Europea en la regulación de la responsabilidad civil por los daños causados por la inteligencia artificial», *Revista española de derecho europeo*, núm. 66, 2018, p. 10, aunque es evidente que las nuevas tecnologías son positivas para el ser humano ya que racionalizan y suprimen laboriosos y arriesgados trabajos, y rompen, tanto en el ámbito laboral como social, las barreras del espacio y el tiempo, la evolución conlleva el surgimiento de «nuevas formas de lesionar los derechos, nuevas formas de dañar al ser humano y por qué no, posibles nuevos derechos que deben protegerse». En sentido similar, entre otros muchos autores, SÁNCHEZ CAPARRÓS, M., *op.cit.*, pp.1-2/34 o GARCÍA GARCÍA, S., «Una aproximación a la futura regulación de la inteligencia artificial en la Unión Europea», *Revista de Estudios Europeos*, N.º. 79, 2022. Para SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 123, a pesar de que es necesario conocer los riesgos que implican estos sistemas para que el ordenamiento jurídico pueda hacerles frente de forma rápida, *no se deben adoptar actitudes tecnófobas* porque son muchos e importantes los efectos positivos de la automatización de decisiones.

¹⁸⁴ Los beneficios potenciales de la IA en numerosos campos como la salud, educación medioambiente, etc., son reconocidos tanto en la legislación europea como en estrategias de digitalización nacionales, lo que justifica su desarrollo y uso. Se ha puesto de manifiesto que la IA puede desempeñar un papel crucial en el cumplimiento de los Objetivos de Desarrollo Sostenible (ODS) de la ONU, tales como la mejora de la salud y la educación, la gestión del cambio climático y la reducción de la pobreza.

algunos de los sistemas de IA de mayor repercusión en el ámbito de la salud, como el de *IBM Watson Health* y *Google DeepMind* que se utilizan en hospitales para analizar grandes cantidades de datos médicos, ayudando a los profesionales a diagnosticar enfermedades como el cáncer, o los robots quirúrgicos asistidos por IA, como el sistema *da Vinci*, que permiten a los cirujanos realizar operaciones complejas con precisión milimétrica, reduciendo el riesgo de errores humanos y mejorando los tiempos de recuperación de los pacientes.

Por ello, debemos ser capaces de aprovechar todas sus posibilidades teniendo muy presentes dos ideas importantes: la primera, que ni ésta ni ninguna otra tecnología que pueda desarrollarse en el futuro serán la panacea que resolverá de forma cuasi milagrosa todos nuestros problemas. Por lo que se refiere al problema concreto analizado en este trabajo, esto es, la toma de decisiones discriminatorias, la IA no tiene tanto poder como para conseguir que todas las decisiones que se adopten mediante estos sistemas sean siempre justas y libres de sesgos; la segunda, que cualquier tecnología emergente implica riesgos y la IA no es la excepción. Ahora bien, en la medida en que algunos de estos peligros pueden pasar desapercibidos para la sociedad (como sucede con la discriminación algorítmica) hay que procurar poner todos los medios disponibles para detectarlos y advertir sobre ellos, con el fin de poder evitarlos, o cuanto menos, minimizarlos.

Con este fin creemos conveniente relatar, siquiera brevemente y sin ánimo de ser exhaustivos, algunos de los casos de discriminación algorítmica que mayor repercusión han tenido en los últimos años. El objetivo principal de dicha exposición es que el lector pueda hacerse una idea de las dimensiones que está alcanzando en la actualidad el fenómeno de la discriminación algorítmica y se percate de que en la práctica estos casos se dan con mayor frecuencia de la que sería deseable. Y es que, creemos junto con GINÉS I FABREGAS, que «*es importante deconstruir el discurso de los algoritmos como modelos estadísticos puramente neutros y matemáticamente objetivos y aceptar la realidad sesgada y discriminatoria de la que se han nutrido*»¹⁸⁵.

La buena noticia es que, como también se analizará en su momento, es posible corregir el sesgo algorítmico para evitar la discriminación injusta¹⁸⁶. De hecho, se ha comprobado que los métodos existentes para hacerlo consiguen reducir considerablemente el porcentaje de error de estos sistemas. Prueba de ello es que, tras la investigación realizada sobre el *software* de reco-

¹⁸⁵ GINÉS I FABREGAS, *op.cit.*, p. 321. Como destaca O'NEIL, *Sesgo codificado*, 2020, la estructura matemática del algoritmo no es racista ni machista, pero los datos que se utilizan para entrenarla reflejan el pasado. Por ello, la máquina lo que hace es reflejar el mundo tal como es. Una máquina no toma decisiones éticas sino sólo decisiones matemáticas.

¹⁸⁶ Nos remitimos al Capítulo IV.

nocimiento facial empleado por IBM que funcionaba peor con mujeres negras que con hombres blancos al haber sido entrenado con una cantidad de imágenes mucho mayor de los segundos que de las primeras, se evidenció que era posible hacer frente a la discriminación algorítmica siempre que hubiera voluntad empresarial, ya que las soluciones técnicas disponibles para ello eran viables.

1.2. CASOS QUE EVIDENCIAN LA DISCRIMINACIÓN ALGORÍTMICA A GRUPOS VULNERABLES

Los sesgos en la IA son tan antiguos como la tecnología misma. El primer caso de sesgo discriminatorio detectado en un sistema de IA aconteció en los años 70¹⁸⁷. Se trató de un programa de la Escuela de Medicina del Hospital St. George para automatizar la lectura de los *currículums* en los procesos de solicitud de admisión. En 1986, se detectó que el programa excluía sistemáticamente a extranjeros y mujeres ocasionando una grave falta de diversidad entre los alumnos aceptados. Años después, la Comisión para la Igualdad Racial de Reino Unido declaró culpable de discriminación a esta institución. El motivo de que el algoritmo discriminara fue que tenía en cuenta ciertas características de las personas como el nombre o el lugar de nacimiento a la hora de priorizar unos candidatos sobre otros.

Son cuantiosos los casos de sesgo algorítmico discriminatorio que han acaecido en los últimos años, tanto en el sector público como en el privado, y que han perjudicado a grupos sociales especialmente vulnerables¹⁸⁸.

¹⁸⁷ Lo recoge DÍAZ ALABART, S., en *Robots y responsabilidad civil*, Madrid, 2018, p. 17, nota 27.

¹⁸⁸ Numerosos estudios muestran cómo decisiones algorítmicas pueden perpetuar estereotipos sociales y provocar discriminación injusta. Uno de los primeros reflejaba cómo los resultados de búsquedas en Google de nombres generalmente asociados a personas negras aparecían junto a anuncios que sugerían detenciones policiales entre el 92-95% de las veces. Vid., SWEENEY, L., «Discrimination in Online Ad Delivery», 2013. Disponible en <http://dx.doi.org/10.2139/ssrn.2208240>. Menciona MUÑOZ RUIZ, A.B., un interesante caso relativo a la Agencia de Empleo austriaca que desde el 2016 evalúa las oportunidades de empleo de los candidatos a un puesto de trabajo con apoyo en un algoritmo que concede menor puntuación a las mujeres, a las personas discapacitadas y a los mayores de 30 años. Vid., «¿Se deben regular los algoritmos? Un breve análisis a la propuesta normativa alemana: la pirámide de criticidad basada en el riesgo», recuperado de <https://www.elforodelabos.es/se-deben-regular-los-algoritmos-un-breve-analisis-a-la-propuesta-normativa-alemana-la-piramide-de-criticidad-basada-en-el-riesgo/>

Se denuncia esta situación en numerosos trabajos doctrinales. Puede verse, entre otros muchos, EUBANKS, V., *op.cit.*, u O'NEIL, *op.cit.* También MARTÍNEZ, N/ MATUTE, H., «Discriminación racial en la Inteligencia Artificial», <https://ethic.es/2020/09/discriminacion-racial-en-la-inteligencia-artificial-sesgos/>; MUÑOZ GUTIÉRREZ, C., *op.cit.*, p. 273 o SÁNCHEZ CAPARRÓS, M., *op.cit.*, p. 17/34; SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 91;

Algunos han ocurrido en grandes multinacionales como Amazon, Apple, IBM o Google. Así, en 2018, se comprobó que el algoritmo usado por Amazon para seleccionar candidatos para un puesto de trabajo era misógino, ya que los resultados que ofrecía presentaban sesgos contra las mujeres. En concreto, a la hora de seleccionar los mejores currículums para trabajos de desarrollo de *software* y otras áreas técnicas, el sistema no estaba aplicando principios de igualdad y meritocracia, sino que valoraba peor a las candidatas mujeres que a los hombres. La razón residía en que el *software* había sido entrenado con datos de las contrataciones llevadas a cabo en la empresa en los últimos 10 años, donde la mayoría de personas seleccionadas para esos puestos habían sido hombres. Dichos datos se ofrecieron al algoritmo sin indicaciones de lo que estaba bien o mal, o de la igualdad de oportunidades que debía buscar, por lo que el *software* interpretó que lo que venía haciéndose en la empresa habitualmente desde hacía años era lo más correcto. Es decir, el algoritmo aprendió que los hombres encajaban mejor en la empresa que las mujeres, perpetuando así la discriminación por razón de género. Lo más curioso, como explicaremos más adelante, es que la variable indicativa del «sexo» no se había introducido en el sistema, pero el algoritmo la infirió de otros datos.

En 2019, el algoritmo para la concesión de crédito de la *Apple Card* fue tildado de ser discriminatorio por razón de género, en concreto, por conceder a un hombre veinte veces más crédito que a su mujer, estando ambos en la misma situación financiera y siendo idéntico su perfil de riesgo¹⁸⁹.

En 2020, se demostró que el *software* para el reconocimiento facial desarrollado por IBM fallaba considerablemente cuando trataba de reconocer a mujeres negras en comparación a lo que ocurría con los hombres caucásicos. El estudio, iniciado años antes, se realizó por el equipo de Joy Buolamwini, investigadora del MIT Media Lab, donde se analizaron tres softwares de reconocimiento distintos: el de Microsoft, el de IBM y Face ++ de Megvii. En concreto, respecto al de IBM se detectó un margen de error del 1% cuando se reconocían fotografías de hombres blancos, margen que aumentaba para el género femenino alcanzando hasta un 7% en mujeres blancas. Mucho más notable fue lo sucedido cuando se varió la raza de los sujetos a reconocer. En el caso de hombres de piel oscura, la tasa de error aumentaba hasta un 12%, porcentaje que se disparó en el caso de las mujeres donde los sistemas fallaron un 35% de ocasiones. En definitiva, mientras que la tasa de error era del 1%

RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», *cit.*, p. 437, cita diversas investigaciones que reflejan cómo personas que ya parten de una situación de desventaja pueden resultar más perjudicadas todavía como consecuencia del uso de algoritmos.

¹⁸⁹ Vid., <https://www.xataka.com/empresas-y-economia/apple-card-sexismo-polemica-algoritmo-apple-concede-a-hombre-20-veces-credito-su-mujer>.

en hombres blancos, ascendía al 35% en mujeres negras¹⁹⁰. La causa de la discriminación residía en que se había entrenado al algoritmo con un número de imágenes de hombres blancos mucho mayor que de otra raza y género, lo que provocó que el sistema tuviera muchos más problemas para reconocerlas.

Otros ejemplos de discriminación algorítmica los protagonizaron Google o Microsoft. En el primer caso, cuando el algoritmo de clasificación de *Google Photos* después de su aprendizaje automático confundió a varios jóvenes afroamericanos con gorilas o cuando se demostró que su algoritmo publicitario mostraba ofertas de empleo mejor remuneradas a los hombres que a las mujeres¹⁹¹. En caso de Microsoft, cuando el *chatbot Tay* que se servía de la IA para mantener conversaciones en una conocida red social con un público de entre 18 y 24 años, empezó a elaborar mensajes de contenido racista, sexista y xenófobo, y tuvo que ser retirado por la compañía al día siguiente de su lanzamiento.

También en el sector de la conducción autónoma se ha observado el sesgo racial, al demostrar un estudio llevado a cabo por investigadores del Instituto Tecnológico de Georgia (EEUU) sobre ocho sistemas de IA empleados en la detección de objetos, que tendían a favorecer los tonos claros de piel sobre los oscuros y discriminaban a los peatones en función de dicho rasgo. Incluso cuando los investigadores matizaron que había imágenes tomadas durante el día y otras por la noche, la IA continuó mostrando una desventaja ante los peatones de tez oscura, lo que llevó a los científicos a concluir que había muchas más probabilidades de que las víctimas de un accidente con un coche autónomo fueran negras que blancas¹⁹².

¹⁹⁰ BUOLAMWINI, J./GEBRU, T., «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender classification», *Proceedings of Machine Learning Research* 81, 2018, pp. 1-15, recuperado de <https://www.xatakamovil.com/movil-y-sociedad/la-ia-de-los-sistemas-de-reconocimiento-facial-tienen-un-problema-racial-discriminacion-de-razas-por-falta-de-muestras>. También hay investigaciones sobre sistemas de reconocimiento facial que fallan a la hora de identificar a personas asiáticas, mostrando mensajes como «¿Parpadeaste?» (vid., ROSE, A., «Are Face-Detection Cameras Racist?», <https://time.com/archive/6906847/are-face-detection-cameras-racist/>, consultado el 21 de diciembre de 2024). También se han dado casos de problemas de los sistemas de IA para el reconocimiento biométrico de personas transexuales. Sobre el tema, vid., <https://www.cnn.com/2018/08/08/transgender-uber-driver-suspended-tech-oversight-facial-recognition.html> y SCHEUERMAN/PAU/BRUBAKER, «How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services», *Proc. ACM Hum. Comput. Interact.* 3, CSCW, Article 144, 2019, <https://doi.org/10.1145/3359246>

¹⁹¹ En 2015, un estudio de la Universidad de Carnegie Mellon demostró que el número de mujeres recibiendo anuncios en Google Ads con ofertas de empleo de más de 200.000 dólares brutos al año era significativamente menor que el de hombres. Vid., <https://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html>

¹⁹² WILSON, B./ HOFFMAN, J./MORGENSTERN, J., «Predictive Inequity in Object Detection», 2019 <https://arxiv.org/pdf/1902.11097.pdf>

En otro estudio realizado respecto de un algoritmo utilizado por aseguradoras y hospitales en EEUU para analizar los riesgos de salud de los pacientes y asignar atención médica, también se advirtió que el sistema discriminaba sistemáticamente a la población afroamericana¹⁹³. Lo curioso fue que, como ocurrió en el caso de Amazon, el algoritmo ni siquiera sabía la raza de los pacientes puesto que ese dato no se le había proporcionado. Como explicaremos al analizar las posibles causas o factores de la discriminación algorítmica, el problema residía en el mal diseño del sistema.

Mucha polémica suscitó también el caso sucedido en Reino Unido en el que, tras la suspensión de los exámenes externos de acceso a la Universidad a consecuencia de la COVID-19 (curso 19-20), se implementó un algoritmo para determinar las notas finales de los estudiantes de bachiller que se tendrían en cuenta para la elección de centro educativo (*Ofqual*)¹⁹⁴. Lo que se pretendía averiguar con este sistema era qué calificación hubieran obtenido esos estudiantes si se hubieran presentado a los exámenes. Para ello, el programa tenía en cuenta las calificaciones existentes de años anteriores y el historial académico de los alumnos –es decir, datos pasados relacionados con la trayectoria del propio centro y con los alumnos particulares– y hallaba correlaciones para predecir la «supuesta» nota de los alumnos que no se habían examinado. Lo que ocurrió fue que el sistema terminó por penalizar a estudiantes de alto rendimiento que residían en barrios desfavorecidos y de bajos ingresos, que normalmente acudían a escuelas públicas cuyas notas medias eran, en general, más bajas que las obtenidas por estudiantes de escuelas privadas con mayores recursos. Es decir, ciertos alumnos se vieron perjudicados por el sistema (y otros beneficiados) por circunstancias ajenas a su persona –como la trayectoria anterior de su centro educativo–. Y es que, como relata HUERGO LORA, «a igualdad de notas, el algoritmo les daba más valor si provenían de un centro que en años anteriores había obtenido mejores notas en los exámenes externos. Como, además, sucede que existe una correlación social (porque, con frecuencia, los centros que obtienen mejores notas en los exámenes externos son centros privados o que tienen un alumnado que procede de entornos de renta más alta o con padres con mayor nivel de estudios), enseguida surgió la acusación de que el algoritmo tenía un sesgo social y finalmente no fue utilizado (en la práctica, cada alumno recibió la calificación más alta comparando la resultante del algoritmo y las notas de su centro educativo)»¹⁹⁵.

¹⁹³ El estudio puede consultarse en <https://www.science.org/doi/abs/10.1126/science.aax2342#con1>

¹⁹⁴ Más información sobre este asunto puede verse en: <https://www.technologyreview.es/s/12573/el-algoritmo-escolar-que-destroza-toda-una-generacion-de-estudiantes>

¹⁹⁵ HUERGO LORA, A., «Administraciones Públicas e inteligencia artificial: ¿más o menos discrecionalidad?», *El Cronista del Estado Social y Democrático de Derecho*, n° 96-97 (Octubre-Noviembre), 2021, p. 5/20. Accesible en <https://laadministracionaldia.inap.es/noticia.asp?id=1512494> (última con-

Junto a los múltiples casos de sesgos raciales y de género que se han ido exponiendo, debe advertirse que la discriminación algorítmica por otros motivos, como la discapacidad o la edad, está ganando terreno a pasos agigantados. Algunos de los ámbitos donde dicho fenómeno amenaza en mayor medida a estas personas son el acceso al empleo o a bienes y servicios¹⁹⁶. En muchas ocasiones esta discriminación se genera a través de sistemas de reconocimiento de datos biométricos cuyo uso es cada vez más frecuente para diversos cometidos (por ejemplo, fichar a la entrada y salida del trabajo con huella dactilar, acceder a determinados dispositivos o aplicaciones mediante el reconocimiento facial o de voz, etc.)¹⁹⁷. Hace algún tiempo se puso de manifiesto que determinados procesos de selección o contratación laboral basados en videojuegos, o que realizaban entrevistas por videoconferencia donde se empleaba un *software* para analizar las expresiones faciales y corporales, el tono o el lenguaje de un candidato, perjudicaban a personas con discapacidad visual o auditiva¹⁹⁸. También se ha denunciado discriminación a través

sulta 7 de noviembre de 2024). Explica el autor que, en todos estos casos, lo que se trata de hallar es un dato desconocido (qué nota habrían obtenido los estudiantes en un examen no realizado) a partir de datos conocidos (en este caso, toda la trayectoria académica del alumno y de los alumnos de años anteriores: notas del centro educativo, trayectoria del propio centro, etc.).

¹⁹⁶ Un problema importante que se ha observado respecto a la discriminación de las personas mayores y/o con discapacidad en el ámbito del acceso a bienes o servicios, es que la misma cae en las grietas del Derecho antidiscriminatorio. Es decir, las Directiva de igualdad en el acceso a bienes y servicios no incluyen la discapacidad o la edad como causas de discriminación. Y es que, en general, el Derecho europeo antidiscriminatorio es fragmentario y parcial. Sobre este problema, en las conclusiones al capítulo 2 del *Estudio sobre discriminación algorítmica en Europa*, ya citado, sus autores concluyen que el uso creciente de algoritmos en todos los ámbitos de la sociedad agudiza este problema y multiplica el riesgo de que la discriminación basada en ciertos motivos como la edad, discapacidad, orientación sexual y religión en el ámbito de los bienes y servicios, donde los anunciantes y los proveedores utilizan cada vez más los algoritmos de elaboración de perfiles, no esté contemplada en el marco legal europeo. Critican estos autores que, en general, el limitado alcance material de la legislación europea reduce su eficacia para erradicar la discriminación algorítmica y observan que, en la medida en que las normas nacionales de igualdad tengan un ámbito de aplicación más amplio y no implementen las posibles excepciones previstas en las Directivas, serán más útiles para prevenir y reparar la discriminación algorítmica. GERARDS, J./XENIDIS, R., *op.cit.*, pp. 75 y ss.

¹⁹⁷ Para RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», *cit.*, p. 432, cuando se utilizan herramientas IA de «screening» u observación en vivo de trabajadores para evaluarlos con fines de contratación o para analizar su rendimiento a efectos de progresión profesional, se puede incurrir en discriminación directa, ya que están diseñadas para la búsqueda de un perfil determinado, asociado a características que incluyen rasgos de la personalidad o elementos asociados al «aspectismo», que pueden servir para descartar del proceso de contratación a ciertas personas.

¹⁹⁸ Sobre estos *softwares* puede verse ampliamente, RAMÍREZ-BUSTAMANTE, N./ PÁEZ, A., «Análisis jurídico de la discriminación algorítmica en los procesos de selección laboral» (2021). Disponible en <https://ssrn.com/abstract=3765741> o <http://dx.doi.org/10.2139/ssrn.3765741>, p. 3, donde se explica cómo funciona uno de los programas de evaluación más sofisticados utilizado por la compañía HireVue4 que analiza los patrones de habla y las expresiones faciales de los aspirantes

de los sistemas de IA integrados en televisores inteligentes o aplicaciones móviles, cuyos asistentes de voz no reconocen la dicción de algunas personas al haber sido diseñados y entrenados con patrones biométricos «normalizados» (problema que está afectando particularmente a las personas con Síndrome de Down). Asimismo, las aplicaciones que usan sistemas de reconocimiento de huella dactilar fallan con los adultos mayores al no tener en consideración que la edad perjudica el dactilograma¹⁹⁹. Es por ello, que respecto de estos colectivos los sistemas de IA en lugar de generar inclusión, pueden acabar excluyendo, y por ende discriminando, bien porque el sistema inteligente no está diseñado para permitir su accesibilidad universal, bien porque, aun estándolo, el programa tiene en cuenta datos o atributos personales como la discapacidad o la edad como variable o criterio «negativo» que acaba excluyendo a la persona mayor o con discapacidad de determinados procesos.

Tanto el uso de sistemas de IA destinados a la contratación o selección de personas físicas, como aquéllos que se utilicen para detectar emociones o determinar la asociación de personas a categorías sociales concretas a partir de datos biométricos²⁰⁰, han sido objeto de especial atención en el RIA por su

en videos grabados mientras responden una serie de preguntas. En p. 14 señalan los autores que algunos académicos han llamado a la selección de personal a través de algoritmos «culling systems», o sistemas de sacrificio, en los que se utiliza la pertenencia de un candidato a una clase (mujeres, ex-convictos, personas mayores, etc.) como información que implica el automático de esos candidatos. Según O'NEIL, estos sistemas puede penalizar el nerviosismo de una persona durante la entrevista o errores lingüísticos de una persona no nativa, *op.cit.*, p. 116. Sobre este tipo de *softwares* también puede verse GINÉS i FABRELLAS, A., *op.cit.*, pp. 296 y ss. Esta autora, sobre la base de diversas investigaciones, afirma que cada vez es más común la utilización de softwares de evaluación de personas (*Candidate Assessment Softwares*—CAS—) que utilizan test de personalidad, inteligencia y/o salud mental para evaluar a las personas candidatas a un puesto de trabajo. En atención a estas características, el sistema elabora un perfil profesional de cada persona, haciendo predicciones sobre sus competencias laborales, comportamiento, capacidad para aprender, responsabilidad y estabilidad personal. La finalidad final del sistema es clasificar a las personas candidatas y ranquearlas en relación con otras en base a su empleabilidad y encaje en el puesto de trabajo ofertado. También menciona otros sistemas de IA para la selección y contratación de personal, por ejemplo, los *bots* para chatear con los candidatos en el proceso de selección, o sistemas de seguimiento de personas candidatas (*Applicant Tracking System*—ATS—), que permiten comparar currículos y analizar actividad en redes sociales para hacer predicciones sobre la adecuación de las personas al puesto de trabajo. También hace referencia a estos sistemas, ALAMEDA CASTILLO, M.T., *op.cit.*, pp. 19-20.

¹⁹⁹ La AEPD, en su *Informe sobre la Adecuación del RGPD a la IA*, cit., p. 37, pone de manifiesto la acusada inexactitud de los sistemas de tratamiento de datos basados en información biométrica, como IA sobre reconocimiento facial, huellas dactilares, voz, etc. Señala que en estos casos se ha de tener en cuenta el impacto sobre la recogida de los datos de personas con alguna discapacidad o singularidad física y que se debe evitar la discriminación por no ser «biométricamente adecuado».

²⁰⁰ Los sistemas de categorización biométrica que clasifiquen individualmente a las personas físicas sobre la base de sus datos biométricos para deducir o inferir su raza, opiniones políticas, afiliación sindical, convicciones religiosas o filosóficas, vida sexual u orientación sexual, están expresa-

potencialidad de afectar a derechos fundamentales, y en particular, a la igualdad y no discriminación.

En particular, sobre los *sistemas de reconocimiento de emociones*, el art. 3.39 los define como aquellos destinados a *distinguir o inferir las emociones o las intenciones de las personas físicas a partir de sus datos biométricos*. El Considerando 18 matiza que se refieren a «emociones o intenciones como la felicidad, la tristeza, la indignación, la sorpresa, el asco, el apuro, el entusiasmo, la vergüenza, el desprecio, la satisfacción y la diversión», pero no incluyen «los estados físicos, como el dolor o el cansancio, por ejemplo, para sistemas utilizados para detectar el cansancio de los pilotos o conductores profesionales con el fin de evitar accidentes. Tampoco incluye la mera detección de expresiones, gestos o movimientos que resulten obvios, salvo que se utilicen para distinguir o deducir emociones. Esas expresiones pueden ser expresiones faciales básicas, como un ceño fruncido o una sonrisa; gestos como el movimiento de las manos, los brazos o la cabeza, o características de la voz de una persona, como una voz alzada o un susurro».

Se reconoce también en el Considerando 44 que «*existe una gran preocupación respecto a la base científica de los sistemas de IA que procuran detectar o deducir las emociones, especialmente porque la expresión de las emociones varía de forma considerable entre culturas y situaciones, e incluso en una misma persona*²⁰¹. Algunas de las deficiencias principales de estos sistemas son la fiabilidad limitada, la falta de especificidad y la limitada posibilidad de generalizar. Por consiguiente, los sistemas de IA que detectan o deducen las emociones o las intenciones de las personas físicas a partir de sus datos biométricos pueden tener resultados discriminatorios y pueden invadir los derechos y las libertades de las personas afectadas. Teniendo en cuenta el desequilibrio de poder en el contexto laboral o educativo, unido al carácter intrusivo de estos sistemas, dichos

mente prohibidos en el art. 5.1.g). Esta prohibición no abarca el etiquetado o filtrado de conjuntos de datos biométricos adquiridos legalmente, como imágenes, basado en datos biométricos ni la categorización de datos biométricos en el ámbito de la aplicación de la ley. Un ejemplo de este tipo de herramientas de IA que proporciona HERRERA DE LAS HERAS, R., «Protección de datos e inteligencia artificial», *cit.*, pp. 665-666, sería el que desarrolló el MIT durante la pandemia de la COVID-19 y que puso a disposición de las autoridades, que permitía distinguir a las personas asintomáticas que estaban infectadas por el virus a través de la grabación de sus toses. La aplicación no identificaba a las personas que tosían, sino que solo usaba sus datos para informar de la posibilidad de que esa tos pudiera implicar tener el virus. Por tanto, en la medida en que el uso de esos datos personales biométricos no permitían identificar a la persona que los cedía (esto es, se trataba de datos anonimizados) no había vulneración del derecho a la privacidad.

²⁰¹ Sobre este tema, destaca GINÉS i FABRELLAS, A., *op.cit.*, p. 302, la exhaustiva investigación de FELDMAN y otros en la que se concluye que, «si bien existen movimientos faciales generalmente asociados a las principales emociones –enfado, asco, miedo, felicidad, tristeza o sorpresa–, existen diferencias entre culturas, situaciones o personas. Los movimientos faciales no pueden automáticamente atribuirse a un concreto e incuestionable estado de ánimo o emoción, existiendo diferencias culturales, contextuales e, incluso, en una misma persona». El estudio citado puede consultarse en <https://pmc.ncbi.nlm.nih.gov/articles/PMC6640856/>

sistemas podrían dar lugar a un trato perjudicial o desfavorable de determinadas personas físicas o colectivos enteros. Por tanto, debe prohibirse la introducción en el mercado, la puesta en servicio y el uso de sistemas de IA destinados a ser utilizados para detectar el estado emocional de las personas en situaciones relacionadas con el lugar de trabajo y el ámbito educativo. Dicha prohibición no debe aplicarse a los sistemas de IA introducidos en el mercado estrictamente con fines médicos o de seguridad, como los sistemas destinados a un uso terapéutico»²⁰².

En los casos en que dichos sistemas no estén prohibidos, será obligatorio notificar a las personas físicas que están expuestas a sistemas que están tratando sus datos biométricos para determinar o inferir sus *emociones o intenciones, o para incluirlas en categorías específicas*²⁰³. Es el art. 50 RIA el que impone una serie de obligaciones de *transparencia a proveedores y responsables del despliegue que serán analizadas en su momento*²⁰⁴.

1.2.1. Particular referencia a tres casos de discriminación algorítmica de gran repercusión que han llegado a los tribunales

Sin perjuicio de todos los casos comentados, que representan un ínfimo porcentaje de los existentes, tres asuntos de discriminación algorítmica requieren especial atención en este momento: el caso del algoritmo *Compas*, el

²⁰² Esta prohibición se contempla en el art. 5.1.f), precepto que recoge las prácticas de IA prohibidas. Así, dispone dicho apartado que «la introducción en el mercado, la puesta en servicio para este fin específico o el uso de sistemas de IA para inferir las emociones de una persona física en los lugares de trabajo y en los centros educativos, excepto cuando el sistema de IA esté destinado a ser instalado o introducido en el mercado por motivos médicos o de seguridad».

²⁰³ Vid., Considerando 132. *Estas categorías específicas pueden referirse a aspectos como el sexo, la edad, el color del pelo, el color de ojos, los tatuajes, los rasgos personales, el origen étnico o las preferencias e intereses personales. Esta información y estas notificaciones deben facilitarse en formatos accesibles a las personas con discapacidad.*

²⁰⁴ Explica ÁLVAREZ CUESTA, H. *op.cit.*, p. 410, que la Ley de Entrevistas de Video de Inteligencia Artificial (*Artificial Intelligence Video Interview Act*) de Illinois (EEUU) regula el uso de sistemas de IA en entrevistas de video para la contratación de personal y establece determinadas exigencias para la implementación de estos sistemas de IA relacionadas con la transparencia: por ejemplo, los empleadores deben notificar a los candidatos cuándo se usa IA en las entrevistas de video, deben explicar cómo funciona el sistema y qué características tiene en cuenta para evaluar la aptitud del candidato, el candidato debe consentir la evaluación por el sistema antes de que comience la entrevista grabada (esto es, se exige un consentimiento informado del solicitante de empleo). Además, la ley estipula que el acceso a esos videos ha de restringirse y que solo podrán acceder a ellos las personas que vayan a evaluar la idoneidad del candidato. También establece que, si el candidato lo solicita, han de destruirse las copias del video y cualquier análisis o resultado derivado de éste dentro de los 30 días siguientes a su solicitud. En definitiva, lo que se pretende con esta Ley es proteger la privacidad de los candidatos y asegurar transparencia en el uso de la IA en procesos de selección de personal, mitigando el riesgo de discriminación algorítmica.

del algoritmo *SyRI* y el del algoritmo *Frank*. Todos ellos han sido objeto de pronunciamientos judiciales pioneros en la materia a los que haremos referencia en distintas partes de esta obra. Por ello, basta por ahora con relatar brevemente los hechos acaecidos y destacar las cuestiones claves para su resolución. Es importante poner de relieve que en los dos primeros asuntos el uso de algoritmos para la toma de decisiones se llevó a cabo por organismos públicos, a diferencia del último en que se utilizó por una empresa privada.

1.2.1.1. *Compas*

El primer asunto en el que nos detendremos es el de *Loomis v. Wisconsin*, resuelto por el Tribunal Supremo de Wisconsin (EEUU), que giró en torno a un sistema inteligente de evaluación de riesgos denominado *Compas* (*Correctional Offender Management Profiling for Alternative Sanctions*)²⁰⁵. El mismo era propiedad de una empresa privada (*Equivant*) y se utilizaba en la jurisdicción penal norteamericana para determinar la extensión y duración de las penas a imponer y el tipo de medidas preventivas a adoptar, además de para definir el tratamiento, supervisión y libertad condicional en la fase penitenciaria²⁰⁶. La herramienta evaluaba la probabilidad de reincidencia de los presos

²⁰⁵ Caso *Loomis vs. Wisconsin*, Decisión de 26 de junio de 2017. 881 N.W.2d749 (Wis. 2016). Sobre este caso, puede verse el comentario de WASHINGTON, A.L. «How to argue with an algorithm: Lessons from the COMPAS ProPublica Debate», *The Colorado Technology Law Journal*, n° 17, 2018, pp. 131-160. También FERNÁNDEZ ARIENZA, A., «Comentario a la sentencia la Corte de apelaciones en el caso “State of Wisconsin v. Eric L. LOOMIS” del Tribunal Supremo del Estado de Wisconsin», *Derecho Digital e Innovación*, n° 5 (abril-junio), 2020. En el análisis de este caso realizado en *State v. Loomis. Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing* accesible en <https://harvardlawreview.org/print/vol-130/state-v-loomis/> se pone de relieve la paradoja de que los jueces deben ser un control de sesgo en una herramienta de IA diseñada para corregir sus propios sesgos a la hora de tomar decisiones.

²⁰⁶ Explica con detalle cómo se están empleando en la justicia penal estadounidense estos sistemas algorítmicos de evaluación de riesgos, SOLAR CAYÓN, J.I., *op.cit.*, pp. 128 y ss. Como señala el autor, «ante un caso determinado, estas herramientas proporcionan a los funcionarios públicos encargados de tomar una decisión una evaluación del riesgo de reincidencia, cuantificando la probabilidad de que un individuo vuelva a cometer un delito y determinando, en función de dicha probabilidad, su nivel de riesgo (normalmente «riesgo bajo», «riesgo medio o moderado» y «riesgo alto»)». Se basan en modelos estadísticos generados automáticamente mediante técnicas de *machine learning*, que a partir del análisis de grandes volúmenes de datos correspondientes a casos pasados (en concreto el sistema COMPAS decidía a partir de los datos de más de 30.000 casos de todo el país), identifican una serie de correlaciones entre factores personales y sociales y el riesgo de comisión de futuros delitos. Los factores que el sistema tiene en cuenta para hacer sus predicciones son, entre otros, indicadores relativos al historial criminal y de comportamiento antisocial del delincuente, su entorno familiar, relaciones personales y estilo de vida, personalidad y actitudes, nivel académico y ocupación laboral, problemas de drogadicción y salud mental y condiciones de residencia.

y en base a ella los clasificaba en uno de los tres niveles de riesgo posibles («riesgo bajo», «riesgo medio o moderado» y «riesgo alto»). La polémica se generó porque un informe de una organización sin ánimo de lucro llamada *ProPublica*²⁰⁷ detectó que el algoritmo empleaba una fórmula que era propensa a señalar a los detenidos afroamericanos como futuros delincuentes, etiquetándolos erróneamente como de «alto riesgo» casi el doble de veces que a los blancos, e inversamente, etiquetaba equivocadamente a los detenidos blancos como de «bajo riesgo» con más frecuencia que a los negros.

En el caso concreto, un ciudadano afroamericano llamado *Eric Loomis* fue condenado a seis años de cárcel por el robo de un coche y cinco años adicionales bajo un régimen de libertad supervisada, pena que el juez impuso basándose, entre otros criterios, en el riesgo de reincidencia indicado por el sistema *Compas* que lo calificó como de «alto riesgo».

El interesado apeló la sentencia alegando que el uso de un sistema algorítmico para la fijación de su condena vulneraba su derecho constitucional al *debido proceso* por tres motivos: primero, porque no se le había proporcionado información suficiente acerca del funcionamiento del programa, lo que impedía que pudiera recurrir adecuadamente la sentencia cuestionando la validez o fiabilidad de sus evaluaciones; segundo, porque para evaluar el riesgo de reincidencia se habían tenido en cuenta características protegidas de las personas, como el género o la raza, lo que vulneraba su derecho a una sentencia imparcial y libre de sesgos; y tercero, porque la evaluación del riesgo se basaba en datos de grupo y no en el carácter y rasgos únicos del condenado, lo que vulneraba su derecho a una sentencia individualizada²⁰⁸. No obstante, la Corte Suprema de Wisconsin dictaminó que el uso de *Compas* no violaba los derechos procesales de *Loomis* y falló en su contra confirmando la validez del sistema. Se basó, principalmente, en el carácter orientativo que tenían para el juez los resultados del programa.

En este asunto se apuntan tres cuestiones que, a nuestro modo de ver, cobran gran importancia en el análisis de los problemas que plantea el uso de sistemas de IA en la toma de decisiones y que serán tratados a lo largo de esta monografía: por un lado, el sesgo algorítmico, en concreto el que provoca discriminación injusta por basarse en características protegidas de las personas; por otro lado, el riesgo de que a través de estos sistemas que se basan en estadísticas generales se comentan errores que afecten a personas concretas; por

²⁰⁷ ANGWIN, J./LARSON, J./MATTU, S./ KIRCHNER, L., «Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks», *ProPublica*, 23 de mayo de 2016. Disponible en <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

²⁰⁸ SOLAR CAYÓN, J.I., *op.cit.*, p. 139.

último, el problema de la opacidad algorítmica, que a nuestro juicio se erige en el principal escollo para detectar y probar la discriminación.

Dichas debilidades de estos sistemas llevaron al tribunal de Wisconsin a concluir que sus resultados no podían utilizarse como *criterio determinante* para decidir si se condenaba o no al acusado, ni la severidad de la condena, sino sólo como *criterio orientativo* a tener en cuenta por los jueces, junto con otros, para tomar una decisión ajustada a las concretas circunstancias del justiciable.

La cantidad de casos de discriminación algorítmica acaecidos en EEUU llevó a que en 2019 se presentara un proyecto de ley de responsabilidad algorítmica en el Estado de Nueva York con el que se pretendía obligar a las empresas que utilizaran estos sistemas a realizar evaluaciones de impacto y auditorías para evitar resultados discriminatorios. Aunque este proyecto no salió adelante, sirvió para tomar conciencia del problema en ese país. En 2022 se presentó un nuevo proyecto de ley en la Cámara de Representantes que tampoco llegó a aprobarse. También se ha publicado una Orden Ejecutiva presidencial para una inteligencia artificial segura y fiable que establece estándares para garantizar que un desarrollo y uso responsable de esta tecnología, protegiendo la privacidad de los ciudadanos, promoviendo la equidad y los derechos civiles, y fomentando la innovación y la competencia²⁰⁹.

Mucho más avanzada es la situación legislativa en la Unión Europea, donde hace apenas unos meses se aprobó el RIA, cuyo art. 5.1.d) prohíbe los sistemas de evaluación de riesgos que, como Compas, vayan dirigidos a evaluar o predecir la probabilidad de que una persona física cometa una infracción penal basándose únicamente en la elaboración del perfil de esa persona o en la evaluación de los rasgos y características de su personalidad. No obstante, se permiten los sistemas de IA utilizados para apoyar la evaluación humana sobre la implicación de una persona en una actividad delictiva que ya se base en hechos objetivos y verificables directamente relacionados con dicha actividad; por tanto, los sistemas podrán ser utilizados para complementar investigaciones delictivas reales, pero no para hacer predicciones especulativas o tomar decisiones automáticas basadas solo en datos subjetivos como, por ejemplo, el nivel socioeconómico, los antecedentes familiares o las características psicológicas o conductuales de las personas²¹⁰.

²⁰⁹ Puede consultarse en: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/?trk=article-ssr-frontend-pulse_x-social-details_comments-action_comment-text

²¹⁰ Así, no caerían bajo la prohibición del art. 5.1.d) aquellos sistemas de IA utilizados, por ejemplo, para analizar patrones en un caso de fraude financiero ya constatado en base a hechos objetivos y verificables, y generara sugerencias que sirvieran al encargado de tomar una decisión para determinar la posible implicación de una persona en el delito, utilizando pruebas verificables como registros o datos bancarios concretos.

1.2.1.2. SyRi

El segundo de los casos apuntados es el del algoritmo SyRI (*Systeem Risico Indicatie* o *Sistema Indicativo del Riesgo*) que fue resuelto por el Tribunal de lo Civil de la Haya en su sentencia de 5 de febrero de 2020²¹¹. Se trataba de un programa informático creado por el Gobierno holandés que mediante IA procesaba una gran cantidad de datos de las personas procedentes de distintos departamentos administrativos con el fin de identificar posibles casos de fraude a la Seguridad Social y a Hacienda. La sentencia concluyó que el empleo de SyRi vulneraba el art. 8 CEDH (en particular el derecho al respeto a la vida privada y familiar) debido a su falta de transparencia y uso sesgado, pues se utilizaba solo en barrios donde residían personas con rentas bajas o minorías²¹². La relevancia de este pronunciamiento, inédito en Europa, reside en que sienta una postura jurisprudencial muy distinta de la recogida en el caso *Compas*, ya que, mientras el tribunal estadounidense decidió proteger al algoritmo y los derechos de propiedad intelectual sobre el mismo, en Europa se opta por la protección de los derechos fundamentales del ciudadano y se falla en contra del uso de algoritmos predictivos y no transparentes en la toma de decisiones individualizadas²¹³. Concretamente, el tribunal determinó que SyRI no cumplía con el equilibrio necesario entre los beneficios que podía implicar la vigilancia del fraude fiscal y la intrusión en el derecho a la privacidad de las personas, y que la falta de información acerca de cómo el sistema gene-

²¹¹ ECLI:NL:RBDHA:2020:865. Sobre este caso pueden verse los trabajos de LAZCOZ MORATINOS / CASTILLO PARRILLA, «Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: El caso SyRI», *Revista chilena de derecho y tecnología*, vol. 9, n. 1, 2020, pp. 207-225; SUBERBIOLA GARBIZU, I., «Inteligencia artificial, opacidad y motivación de las decisiones administrativas automatizadas», en *La protección de los derechos fundamentales en el ámbito tributario*, coord. por Antonio Vázquez del Rey Villanueva e Irune Suberbiola Garbizu y dir., por Isaac Merino Jara, 2021, pp. 291-328. Recuperado de la Ley Digital (LA LEY 1117/2021); MEUWESE, A., «Regulating algorithmic decision-making one case at the time. Case note on: District Court of The Hague, 5/02/20, ECLI:NL:RBDHA:2020:865 (NJCM vs the Netherlands (SyRI))», *European Review of Digital Administration & Law*, 1(1), 2020, pp. 209-211.

²¹² Como explica ESPÍN ALBA, I., *op.cit.*, pp. 147-148, en este caso fueron varias organizaciones vinculadas a la defensa de los derechos humanos las que demandaron al Estado alegando que los resultados de Syri eran discriminatorios para colectivos históricamente marginados. No obstante, el Estado alegó que la herramienta era acorde con la legislación de protección de datos y la protección de la privacidad, ya que los procedimientos asociados con los que los datos podían vincularse y analizarse se llevaban a cabo de forma anónima y en un entorno seguro.

²¹³ El diferente posicionamiento de los sistemas judiciales europeos y estadounidenses acerca del derecho de acceso a los algoritmos se basa, principalmente, en que en Europa el marco jurídico es más protector de los intereses privados gracias al RGPD, que establece un derecho a la información sobre la lógica subyacente de las decisiones tomadas mediante algoritmos. Así se explica en la *Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno* del Consejo de Europa.

raba sus modelos de riesgo y evaluaba los indicadores socavaba la posibilidad de que los afectados pudieran comprender y cuestionar el procesamiento de sus datos.

Profundizaremos en esta resolución –que establece un precedente significativo al vincular expresamente el derecho a la privacidad con el uso de herramientas de vigilancia basadas en datos– y en los argumentos en que se apoya, cuando analicemos el problema de la opacidad algorítmica.

1.2.1.3. *Frank*

Por último, resta comentar el polémico caso de la empresa *Deliveroo*, que motivó incluso una reforma de la legislación laboral en nuestro país para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales²¹⁴. La citada compañía de comida a domicilio utilizaba un algoritmo conocido como *Frank* para la asignación de franjas horarias a sus trabajadores. Dicho algoritmo tenía en cuenta tanto la fiabilidad como la disponibilidad de los repartidores. La *fiabilidad* se refería al número de veces que el trabajador no atendía a una franja horaria previamente reservada y la *disponibilidad* al número de horas de conexión a la plataforma en horario de alta demanda (básicamente de las 20 a 22 h de viernes a domingo)²¹⁵. Cuantas más horas estuviera disponible el trabajador y menos cancelaciones de franjas horarias previamente reservadas tuviera, mejor posicionamiento le asignaba el sistema en el ranking de la empresa, lo que a su vez implicaba

²¹⁴ Esta reforma fue realizada a través del Real Decreto-ley 9/2021, de 11 de mayo, por el que se modifica el texto refundido de la Ley del Estatuto de los Trabajadores, aprobado por el Real Decreto Legislativo 2/2015, de 23 de octubre, para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales y, posteriormente convalidado por la Ley 12/2021, de 28 de septiembre. Se modifica el art. 64.4 ET para incluir el derecho de la representación legal de la plantilla a ser informada de «los parámetros, reglas e instrucciones en los que se basan los algoritmos o sistemas de inteligencia artificial que afectan a la toma de decisiones que puedan incidir en las condiciones de trabajo, el acceso y mantenimiento del empleo, incluida la elaboración de perfiles». Se establece, así, el derecho de los representantes de los trabajadores a ser informados por la empresa de los parámetros, reglas e instrucciones en los que se basan los algoritmos o sistemas de IA que afecten a la toma de decisiones y que puedan incidir en las condiciones de trabajo, acceso y mantenimiento del empleo, incluida la elaboración de perfiles.

²¹⁵ CASTILLO PARRILLA, J.A., «Sentencia del Tribunal Ordinario de Bolonia de 31 de diciembre de 2020 (Caso Deliveroo) ¿Discriminación algorítmica o discriminación a través de un algoritmo?», *Derecho Digital e Innovación. Digital Law and Innovation Review*, n.7 (octubre-diciembre), 2020, explica que «el índice de fiabilidad tiene en cuenta el número de ocasiones en que el *rider* no cumple con una sesión de trabajo previamente reservada (a través de SSB), entendiéndose por incumplimiento no haber iniciado sesión en la plataforma en los quince primeros minutos de la franja horaria reservada (y encontrándose el *rider* físicamente dentro del área geográfica donde va a repartir)».

que la próxima vez que accediera a la plataforma tuviera preferencia para elegir el horario que más le conviniera para prestar sus servicios²¹⁶. El problema surgió porque el algoritmo no diferenciaba los motivos por los que un *rider* cancelaba una reserva, penalizando igualmente a los que lo hacían por motivos amparados en la legislación laboral, que a los que lo hacían por causas no protegidas jurídicamente²¹⁷.

Tras la demanda planteada por varios sindicatos contra *Deliveroo*, el Tribunal Ordinario de Bolonia concluyó, en una pionera sentencia de 31 de diciembre de 2020²¹⁸, que el algoritmo *Frank* discriminaba indirectamente a ciertos trabajadores en las condiciones de acceso al trabajo porque para llevar a cabo el ranking reputacional de cada repartidor no tenía en cuenta los motivos por los que estos habían tenido que cancelar las sesiones de trabajo previamente reservadas. Según se explica en la sentencia, las instrucciones que la empresa proporcionó al algoritmo ignoraban estos motivos de forma consciente. Por tanto, la programación o diseño del algoritmo no era adecuada cuando penalizaba de la misma forma al que cancelaba su reserva por motivos justificados (ej. enfermedad, ejercicio legítimo del derecho a la huelga...), que a aquellos que lo hacían por motivos no justificados. A juicio del tribunal, se trata de una discriminación *indirecta* ya que a través de una conducta aparentemente neutra (uso de criterios de *fialibilidad* y *disponibilidad*) se perjudica a un grupo protegido. En este caso, se limitaba la posibilidad de acceder a futuras opciones de empleo y retribución a aquellos trabajadores que habían

²¹⁶ Sobre la cada vez más frecuente toma de decisiones algorítmicas por las plataformas digitales de reparto como Glovo, Deliveroo o UberEats, que han sido pioneras en el uso de algoritmos para la toma de decisiones automatizadas en el ámbito laboral puede verse GINÈS i FABRELLAS, A., *op.cit.*, p. 303. Sobre el caso Deliveroo señala la autora que la actividad pasada en la plataforma se utilizaba para determinar la actividad futura, beneficiando a aquellas personas con mayor dedicación y disponibilidad. Con más detalle explica FERNÁNDEZ SÁNCHEZ, S., en «Frank, el algoritmo consciente de Deliveroo. Comentario a la Sentencia del Tribunal de Bolonia 2949/2020, de 31 de diciembre», *Revista de Trabajo y Seguridad Social. CEF* n. 457, pp. 179-193, que «Cada domingo, los/las riders de Deliveroo recibían un mensaje donde se les informaba, en función del ranking, sobre la franja horaria en la que podían acceder a realizar las reservas de las sesiones de trabajo para la semana sucesiva. Así, cada lunes, los/las riders con mejor ranking reputacional accedían a la plataforma para realizar las reservas a las 11:00 de la mañana, los siguientes en el ranking accedían a las 15:00, mientras los que poseían una puntuación peor tenían que esperar hasta las 17:00 horas. Esto comportaba que las ocasiones de trabajo de los/las riders con una puntuación inferior se redujesen según avanzaban las franjas. De tal forma, Frank iba lentamente excluyendo del ciclo productivo a la persona repartidora que no aseguraba la disponibilidad».

²¹⁷ Deliveroo recogía solo dos casos de ausencia al puesto de trabajo: por un lado, accidente del *rider* que le impidiera seguir repartiendo; y por otro lado, que la plataforma dejara de funcionar por problemas técnicos. No obstante, no contemplaba otras causas de ausencia como la enfermedad del trabajador, el cuidado de menores o citaciones judiciales, entre otras.

²¹⁸ Núm.2949/2019, asunto Deliveroo Italia S.R.L.

cambiado de opinión en su reserva horaria y se ausentaban temporalmente de la plataforma por motivos legítimos como el cuidado de menores o la incapacidad temporal, entre otros.

Aunque no se obliga expresamente en la sentencia a que *Deliveroo* publique o modifique el funcionamiento de su algoritmo, se la condena a eliminar los efectos de dicha conducta discriminatoria, eliminación que lógicamente obligaba a modificar ciertos parámetros de puntuación del algoritmo para excluir, como susceptibles de puntuación negativa, incumplimientos justificados de la sesión de trabajo reservada.

Es importante poner de relieve que el tribunal no cuestiona la legitimidad del sistema algorítmico utilizado por la plataforma, ya que entiende que el interés de la empresa en evitar cancelaciones de reservas es legítimo y que atender a la fiabilidad del trabajador para la asignación de franjas horarias tampoco es ilegal. Ahora bien, pone en duda que la imposibilidad de justificar dichas ausencias al puesto de trabajo, penalizando en la elección de futuros trabajos a personas que cancelaron por causas justificadas, se adecúe al Derecho laboral, ya que, al tratar igual a personas que no se encuentran en análoga situación (por cuanto que unas causas de absentismo laboral están amparadas por la legislación y otras no), se las está discriminando injustamente²¹⁹.

A mayor abundamiento, la doctrina ha apuntado con acierto que utilizar la *fiabilidad y disponibilidad* horaria tienen un impacto negativo mucho mayor en el colectivo de las mujeres por cuanto son ellas las que mayoritariamente se acogen a reducciones de jornada y tienen mayores necesidades de conciliación que los hombres. Al no poder conectarse en horas de alta demanda, o tener que cancelar franjas previamente reservadas para dedicarse al cuidado de descendientes o ascendientes en situación de dependencia, ven reducidas sus posibilidades de atender tanto pedidos actuales, como también futuros, al no poder elegir las franjas horarias donde se realizan más pedidos²²⁰.

No es éste un caso aislado en el ámbito laboral sino que, muy al contrario, este sector es campo abonado para la discriminación algorítmica, tanto en el acceso al empleo como en la asignación de tareas, determinación de las

²¹⁹ Como expone ASQUERINO LAMPARERO, «Algoritmo y discriminación», https://www.cielolaboral.com/wp-content/uploads/2022/04/asquerino_noticias_cielo_n4_2022.pdf, no tiene la misma relevancia jurídica la ausencia al trabajo por enfermedad o por el ejercicio del derecho a la huelga, que el absentismo por desidia, y esta diferencia tan asumida por los empresarios no parece ser entendida por Frank.

²²⁰ Así lo explica GINÈS i FABRELLAS, A., *op.cit.*, pp. 305-6, para quien «el interés empresarial de fomentar el trabajo en horas de alta demanda no resulta una causa objetiva, razonable y proporcionada para excluir la discriminación indirecta, al penalizar hacia el futuro –y no solo premiar en el presente– la falta de disponibilidad».

condiciones laborales o incluso extinción de la relación laboral. La doctrina iuslaboralista que ha tratado el asunto expone múltiples ejemplos de lo afirmado que, junto con el famoso caso de la empresa *Deliveroo*, dan buena cuenta de este creciente problema²²¹. Sin ir más lejos, la sentencia de la Audiencia Nacional de 10 de mayo de 2021, aunque *obiter dicta*, detecta discriminación indirecta por razón de edad en un sistema algorítmico de control de la jornada que no permitía registrar pausas para usar el baño²²². Ese algoritmo exigía que el trabajador estuviera continuamente conectado (telefónica y digitalmente) y solo podía realizar las pausas cuando la empresa las tenía programadas o, en su defecto, durante el descanso para comer. Si tenía que interrumpir su actividad para satisfacer «tal necesidad inevitable» se veía obligado, bien a registrar un descanso en un momento no programado por la empresa, bien a no registrarlo y desatender su puesto de trabajo, pudiendo ser sancionado por ello. Para el tribunal, esta práctica aparentemente neutra, además de vulnerar la dignidad de los trabajadores («*que no es otra cosa que el derecho que tiene a ser tratado como una persona en todo momento*») y ser contraria a la protección de la salud, podía constituir «*una discriminación indirecta por razón de edad prosrita por el art. 17.1 del E.T, pues resulta indiscutible que esta práctica, aparentemente neutra, implica en la práctica un trato peyorativo a los trabajadores de más edad respecto de los*

²²¹ Por lo que se refiere al uso de estos sistemas automatizados para despedir trabajadores, puede verse la doctrina citada por GINÈS i FABRELLAS, A., *op.cit.*, p. 297 y «Análítica de personas y discriminación algorítmica en procesos de selección y contratación», *Labos*, Vol. 5, Número extraordinario sobre Normativa europea sobre inteligencia artificial, p. 102, sobre diversos ejemplos como el de la plataforma digital Uber que automáticamente desactiva –eufemismo de la economía de plataformas para referirse al despido– a las personas con una puntuación inferior a 4,5. Otro ejemplo es la empresa Amazon, que rastrea a las personas que trabajan en sus almacenes en Estados Unidos, mide el número y velocidad de cajas empaquetadas por hora y genera avisos automáticos o, incluso, despidе de forma absolutamente automatizada a aquellas personas que no alcanzan la cuota mínima de productividad, penalizando incluso pausas para ir al baño o comer. Y en el ámbito de la selección de personal, cita el caso de GILD, una empresa que elabora para sus clientes (la mayoría empresas tecnológicas) perfiles de candidatos a un puesto de trabajo y que hace unos años decidió incorporar como criterio para la selección de programadores, no sólo su currículum sino su «capital social», que incluía variables como sus contribuciones a plataformas colaborativas como GitHub, el tono de sus tuits o variables *proxy*, los lugares en que quedaban con sus amigos y las webs que visitaban en su tiempo libre, incluida la visita a una página web concreta de manga japonés. El algoritmo infirió que las personas que frecuentaban páginas web de manga japonés tenían mayores habilidades para programar. La discriminación residía en que, como regla general, dichas páginas son visitadas en mayor medida por personas del sexo masculino en su tiempo libre, lo que llevaba a penalizar o descartar del proceso de selección laboral a las mujeres programadoras.

²²² SAN 104/2021 de 10 de mayo de 2021, Rec. 105/2021. Se estima la demanda interpuesta por CCOO contra *Extel Contact Center Sau* a la que se adhieren UGT, SOA y CIG y se declara «el derecho del personal que presta servicios para la demandada al uso del lavabo para atender sus necesidades fisiológicas por el tiempo imprescindible y la correlativa obligación de la empresa de registrar estas pausas de forma separada del resto de descansos y pausas contempladas en el convenio colectivo de Contact Center».

más jóvenes», teniendo en cuenta que es «un hecho notorio que la capacidad de contener tales necesidades a partir de la adolescencia es inversamente proporcional a la edad de las personas»²²³.

Pues bien, todos los asuntos comentados en las páginas anteriores y muchos otros a los que nos referiremos en este trabajo, reflejan la imperiosa necesidad de abordar el tema de la discriminación algorítmica con urgencia y desde un punto de vista interdisciplinar²²⁴.

2. LA DISCRIMINACIÓN ALGORÍTMICA: CONCEPTO, CARACTERÍSTICAS, TIPOS Y CAUSAS

2.1. EL DERECHO A LA IGUALDAD DE TRATO Y LA PROHIBICIÓN DE DISCRIMINACIÓN: CONCEPTO Y RECONOCIMIENTO LEGAL

Es sobradamente conocido por todos, legos o duchos en ciencias jurídicas, que la igualdad de trato y la prohibición de discriminación son derechos fundamentales reconocidos a todos los niveles. Por lo que a nuestro ordenamiento nacional se refiere, es el art. 14 CE el que consagra, por un lado, una declaración general de igualdad ante la ley, y por otro lado, una prohibición de discriminar en base a una serie de motivos²²⁵. Pero además de erigirse como derechos fundamentales (amén de *humanos*), la igualdad y la no discriminación se configuran como principios y mandatos que deben vertebrar

²²³ Vid., FD 4^o. Actualmente se ha puesto en boga en término «edadismo» para hacer referencia a la discriminación por razón de edad. Según la OMS, se denomina «edadismo» a los estereotipos (cómo pensamos), los prejuicios (cómo nos sentimos) y la discriminación (cómo actuamos) dirigidos hacia otras personas o hacia nosotros mismos en función de la edad. *Informe mundial sobre el edadismo*. Washington, D.C. Organización Panamericana de la Salud, 2021, p. 2.

²²⁴ Para BAROCAS, S./SELBST, A. D., «Big Data's Disparate Impact», 104 *California Law Review* 671, 2016, disponible en <https://ssrn.com/abstract=2477899> o <http://dx.doi.org/10.2139/ssrn.2477899>, abordar el impacto discriminatorio de los algoritmos requiere un enfoque interdisciplinario y una transformación en la manera en que las leyes y las políticas públicas entienden la discriminación. Y según mantiene HACKER, P., «Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law», 55 *Common Market Law Review*, 2018, pp. 31-34, la combinación de herramientas legales, enfoques técnicos y colaboración entre disciplinas es esencial para garantizar la equidad y la no discriminación en la era de la IA.

²²⁵ Señala el TC que la virtualidad del art. 14 CE no se agota en la cláusula general de igualdad con la que se inicia su contenido, sino que a continuación el precepto se refiere a la prohibición de una serie de motivos o razones concretos de discriminación que no suponen una lista cerrada de supuestos de discriminación (STC 75/1983, de 3 de agosto, FJ 6), ya que el propio art. 14 CE cierra su enumeración de circunstancias impositivas de discriminación con la referencia genérica a «*cualquier otra condición o circunstancia personal o social*».

la actuación de los poderes públicos a la hora de crear y aplicar las normas jurídicas. En consecuencia, deben estar presentes en cualquier ámbito privado o público en el que las personas puedan ejercer sus derechos, libertades y obligaciones.

Se trata de un principio ético fundamental para alcanzar una convivencia pacífica, que se sustenta en valores personales y sociales que han pasado de ser una declaración de intenciones incluida en los códigos éticos de las organizaciones, a configurar un marco jurídico vinculante de garantías para evitar que las personas no puedan disfrutar de sus derechos en condiciones de igualdad con otras personas, debido a una distinción injustificada que se establece en la política, la ley o el trato aplicados.

Este *principio universal* de igualdad de trato, derivado de la dignidad de las personas, está reconocido internacionalmente y se vulnera cuando se produce cualquier tipo de discriminación injusta. Junto con el principio de no discriminación, son transversales en todo el derecho internacional de derechos humanos. Están consagrados en la Declaración Universal de Derechos Humanos, el Convenio Europeo para la Protección de los Derechos Humanos y de las Libertades Fundamentales, y en otros tratados internacionales específicos como el Convenio de la Organización Internacional del Trabajo n°111, que prohíbe la discriminación en el ámbito del empleo y de la ocupación.

Por lo que se refiere en concreto al ámbito europeo, la protección contra la discriminación puede encontrarse tanto en el Derecho originario de la Unión (Carta de Derechos Fundamentales de la Unión Europea, en adelante CDFUE) como en la normativa del Consejo de Europa (Convenio Europeo de Derechos Humanos, en adelante CEDH) y se asienta en prohibiciones generales de discriminación en base a una serie de categorías protegidas²²⁶. Ha sido objeto de desarrollo específico a través de diversas Directivas de igualdad (denominadas también Directivas antidiscriminatorias) que recogen distintos tipos de discriminación²²⁷.

²²⁶ El art. 2 del Tratado de la Unión Europea (TUE) también reconoce el derecho a la igualdad de trato y la prohibición de discriminación como fundamento de la Unión y el artículo 3.3 se refiere concretamente a la igualdad entre mujeres y hombres. Se reconoce también en los arts. 18-25 del Tratado de Funcionamiento de la Unión Europea (TFUE).

²²⁷ Estas Directivas han ido perfilando, con una sistemática muy parecida, la tutela antidiscriminatoria en Europa. Son las siguientes: Directivas 2000/43/CE, sobre igualdad de trato de las personas independientemente de su origen racial o étnico y la 2000/78/CE, para luchar contra la discriminación por motivos de religión o convicciones, de discapacidad, de edad y de orientación sexual en el ámbito del empleo y en la ocupación y a fin de establecer un marco general para la aplicación del principio de igualdad de trato por los Estados miembros. Por lo que se refiere concretamente a las Directivas sobre igualdad de género: Directiva 2006/54/CE, relativa a la aplicación del principio de igualdad de oportunidades e igualdad de trato entre hombres y mujeres en asuntos de empleo

La CDFUE contiene una lista de derechos humanos inspirada en los derechos reconocidos en las Constituciones de los Estados miembros, el CEDH y los tratados universales de derechos humanos. Por lo que se refiere en concreto a la igualdad, en sus arts. 20 a 26 hace hincapié en la importancia del principio de igualdad de trato en el ordenamiento jurídico de la Unión.

Mientras que el art. 20 contiene el principio general de igualdad en virtud del cual *todas las personas son iguales ante la ley*, que ha sido recogido en todas las constituciones europeas y reconocido por el Tribunal de Justicia como un principio básico del Derecho comunitario, el art. 21 consagra la interdicción de discriminación por una serie de motivos no taxativos; así, se prohíbe *toda* discriminación «y en particular la ejercida por razón de *sexo, raza, color, orígenes étnicos o sociales, características genéticas, lengua, religión o convicciones, opiniones políticas o de cualquier otro tipo, pertenencia a una minoría nacional, patrimonio, nacimiento, discapacidad, edad u orientación sexual*».

Según el TJUE, el *principio de igualdad de trato* es un principio general del Derecho de la Unión, del que el *principio de no discriminación* es una expresión particular. Por su parte, la doctrina viene entendiendo que «igualdad» y «no discriminación» son conceptos diferentes aunque guardan una relación de género (igualdad) a especie (prohibición de discriminación), configurándose la «discriminación» como una forma especial de desigualdad caracterizada por la naturaleza despreciable del prejuicio social descalificatorio, que tiende a tomar como objeto de persecución un rasgo físico o cuasifísico, afectando gravemente a la dignidad de la persona y por consiguiente a la igualdad más básica de los portadores de tal rasgo²²⁸. Y en la misma línea nuestro Alto Tribunal viene afirmado que «lo que caracteriza a la prohibición de discriminación frente al principio genérico de igualdad, es la naturaleza particularmente odiosa del criterio de diferenciación utilizado, que convierte en ele-

y ocupación que compendia las primeras Directivas de los años 70 y 80 del siglo pasado (Directivas 75/117, 76/207/, 97/80 y 86/378/CEE), al tiempo que incorpora las novedades planteadas por la doctrina del TJUE; la Directiva 2004/113/CE, por la que se aplica el principio de igualdad de trato entre hombres y mujeres al acceso a bienes y servicios y su suministro y la 2010/41/UE, sobre la aplicación del principio de igualdad de trato entre hombres y mujeres que ejercen una actividad autónoma. Estas Directivas se completan con la recientemente adoptada Directiva (UE) 2024/1385 del Parlamento Europeo y del Consejo, de 14 de mayo de 2024, sobre la lucha contra la violencia contra las mujeres y la violencia doméstica.

²²⁸ Así lo entiende BARRERO ORTEGA, A., «La tutela antidiscriminatoria en la jurisprudencia del tribunal de justicia de la unión europea», en J.M. MORALES ORTEGA (dir.), *Realidad social y discriminación. Estudios sobre diversidad e inclusión laboral*, Murcia, 2022, p. 44. Señala este autor que «La prohibición de discriminación es una variedad de la igualdad cuando el criterio de desigualdad que concurre es una circunstancia personal especialmente odiosa y, por ende, sospechosa. En teoría, hay que distinguir entre el principio general y la doctrina de la diferencia jurídica de trato razonable y la prohibición de discriminación *stricto sensu*».

mento de segregación, cuando no de persecución, un rasgo o una condición personal innata o una opción elemental que expresa el ejercicio de las libertades más básicas, resultando así un comportamiento radicalmente contrario a la dignidad de la persona y a los derechos inviolables que le son inherentes (art. 10 de la Constitución)»²²⁹.

En el lenguaje ordinario, discriminar es dar un trato desigual y peor a una persona o colectivo que el que se dispensa al común de los individuos en la misma situación, sin que haya una razón lícita para hacerlo²³⁰. Es decir, consiste en tratar de forma diferente y perjudicial a una persona o grupo, siempre que dicho trato se base en ciertas características personales como el sexo, el género, la edad o la raza, sobre todo cuando dichas características no son relevantes en el contexto en que se adopta la decisión. En la medida en que se trata de características que se corresponden con condiciones de la persona que en principio son inmutables (edad, discapacidad, raza, etc), o que están estrechamente relacionadas con su dignidad y autonomía (como las creencias religiosas u opiniones políticas), son objeto de una especial protección legal y se atribuye un mayor desvalor a las decisiones que discriminan en base a ellas. Por tanto, las denominadas «características protegidas», son aquellas condiciones de la persona que no se deben considerar relevantes para administrarle un trato diferenciado u otorgarle un determinado beneficio²³¹. Tanto es así, que no cualquier desigualdad de trato conlleva discriminación prohibida o injusta, sino sólo la basada en tales características o motivos especialmente protegidos²³².

²²⁹ Por todas, la STS 62/2008, de 26 de mayo.

²³⁰ Así la define, por ejemplo, DÍAZ ALABART, S., «La responsabilidad civil por discriminación de las personas por causa de su discapacidad (Art. 18 Liondap)», en J.A. MORENO MARTÍNEZ (coord.), *La responsabilidad civil y su problemática actual*, Madrid, 2007, p. 186. Según el Diccionario de la Real Academia de la Lengua Española, discriminar es «Dar trato desigual a una persona o colectividad por motivos raciales, religiosos, políticos, de sexo, de edad, de condición física o mental, etc.». Sobre el concepto de discriminación, puede verse ampliamente también el *Manual de legislación europea contra la discriminación (edición 2018)*, Luxemburgo, 2019, p. 46.

²³¹ ARCOS VARGAS, M., «La no discriminación en el Derecho derivado de la Unión Europea» en J.M. MORALES ORTEGA (dir.), *Realidad social y discriminación. Estudios sobre diversidad e inclusión laboral*, Murcia, 2022, pp. 26 y ss.

²³² Las Directivas europeas de igualdad prohíben el trato diferenciado basado en ciertas «características protegidas», concretas y limitadas. En particular, por razón de raza o etnia (Directiva sobre la igualdad racial), religión o creencias, orientación sexual, discapacidad y edad (Directiva de igualdad en el empleo), y género (Directiva sobre la igualdad de trato entre hombres y mujeres en el acceso a bienes y servicios, Directiva sobre la igualdad de género (refundición)...). Por el contrario, el CEDH contiene una lista abierta que coincide con las Directivas, pero que va más allá, porque en el art. 14 se establece la prohibición de discriminación por razón de «sexo, raza, color, lengua, religión, opiniones políticas u otras, origen nacional o social, pertenencia a una minoría nacional, fortuna, nacimiento o cualquier otra situación», cláusula de cierre similar a la empleada por nuestro legislador constituyente.

Además, en algunos casos la diferencia de trato puede estar justificada y ser legalmente permisible. Así, aunque los supuestos en los que puede producirse una situación de discriminación son muy variados, no se considerarán tales cuando respondan a una *finalidad legítima y los medios para obtenerla sean adecuados y necesarios*. Es decir, no podría considerarse que existe discriminación si el trato desigual está justificado, justificación que debe apreciarse en relación a la finalidad y efectos de la medida considerada, debiendo darse una relación razonable de proporcionalidad entre los medios empleados y la finalidad perseguida²³³. En definitiva, como viene manteniendo nuestro TC, debe existir una *causa objetiva y razonable* que avale la diferencia de trato para descartar la discriminación²³⁴.

La Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, en coherencia con la doctrina constitucional, consagra esta idea en su art. 4.2 en virtud del cual: «No se considera discriminación la *diferencia de trato* basada en alguna de las causas previstas en el apartado 1 del artículo 2 de esta ley derivada de una disposición, conducta, acto, criterio o práctica *que pueda justificarse objetivamente por una finalidad legítima y como medio adecuado, necesario y proporcionado para alcanzarla*». Y el art. 2.2 dispone que «[...] *podrán establecerse diferencias de trato cuando los criterios para tal diferenciación sean razonables y objetivos*

²³³ Explica BARRERO ORTEGA, A., *op.cit.*, pp. 48-49, que aunque es el autor de la norma el que debe apreciar si se han de contemplar excepciones al trato igualitario, tal valoración encuentra sus límites en la desigualdad irrazonable y la desproporción del tratamiento diferenciado. Por tanto, la clave está en la razonabilidad y proporcionalidad de la desigualdad. Ahí es donde diariamente se mueve la interpretación del juez europeo y ahí es donde hay que acertar. Sobre todo, en la identificación de la causa objetiva y razonable que avala la diferencia de trato, descartando así el trato discriminatorio directo. Entre otras muchas, puede verse la STJUE de 17 de abril de 2018, asunto C-414/16, Vara Egenberger/Evangelisches Werk für Diakonie und Entwicklung eV, ECLI:EU:C:2018:257.

²³⁴ Así se pone de relieve en la jurisprudencia constitucional. Vid., por todas, la STC 61/2013, de 14 de marzo. El TC declara que «el principio de igualdad no implica en todos los casos un tratamiento legal igual con abstracción de cualquier elemento diferenciador de relevancia jurídica, de manera que no toda desigualdad de trato normativo respecto a la regulación de una determinada materia supone una infracción del mandato contenido en el art. 14 CE, sino tan sólo las que introduzcan una diferencia entre situaciones que puedan considerarse iguales, *sin que se ofrezca y posea una justificación objetiva y razonable para ello*, pues, como regla general, el principio de igualdad exige que a iguales supuestos de hecho se apliquen iguales consecuencias jurídicas y, en consecuencia, veda la utilización de elementos de diferenciación que quepa calificar de arbitrarios o carentes de una justificación razonable»; asimismo, «[t]ambién es necesario, para que sea constitucionalmente lícita la diferencia de trato, que las consecuencias jurídicas que se deriven de tal distinción sean proporcionadas a la finalidad perseguida, de suerte que se eviten resultados excesivamente gravosos o desmedidos»; en consecuencia, hemos concluido que «[e]n resumen, el principio de igualdad no sólo exige que la diferencia de trato resulte objetivamente justificada, sino también que supere un juicio de proporcionalidad en sede constitucional sobre la relación existente entre la medida adoptada, el resultado producido y la finalidad pretendida» (STC 63/2011, de 16 de mayo, FJ 3; o SSTC 117/2011, de 4 de julio, FJ 4 y 79/2011, de 6 de junio, FJ 3).

y lo que se persiga es lograr un propósito legítimo o así venga autorizado por norma con rango de ley, o cuando resulten de disposiciones normativas o decisiones generales de las administraciones públicas destinadas a proteger a las personas, o a grupos de población necesitados de acciones específicas para mejorar sus condiciones de vida o favorecer su incorporación al trabajo o a distintos bienes y servicios esenciales y garantizar el ejercicio de sus derechos y libertades en condiciones de igualdad».

2.2. TIPOS DE DISCRIMINACIÓN

2.2.1. Discriminación directa e indirecta

Las Directivas de igualdad suelen diferenciar principalmente entre la discriminación *directa* e *indirecta* y las definen de forma muy similar variando tan solo la causa por la que se discrimina. A tenor de las mismas, existirá «discriminación directa» cuando una persona sea tratada de manera menos favorable de lo que haya sido o vaya a ser tratada otra persona en situación comparable, debido a una característica concreta de dicha persona incluida entre los «motivos protegidos». Ese «trato menos favorable» se determina por comparación entre la supuesta víctima y otra persona que no posea la característica protegida y se encuentre en una situación similar. Por tanto, para determinar si una persona ha sido objeto de un trato menos favorable, es necesario identificar una «referencia comparativa», es decir, una persona que se halle en circunstancias sustancialmente iguales y cuya principal diferencia respecto a la otra sea un motivo protegido²³⁵. Además, es necesario que exista una relación de causalidad entre el trato menos favorable y el motivo protegido, ya que, si la diferencia de trato no se debe a dicho motivo sino a que los sujetos están en circunstancias distintas, no podría considerarse que existe discriminación²³⁶.

También hay que tener en cuenta que, tal como se reconoce en la normativa antidiscriminatoria, puede existir discriminación no sólo cuando se trata de modo diferente a personas en situaciones similares, sino también cuando se trata de forma idéntica a personas en situaciones diferentes²³⁷. Esta última

²³⁵ Excepción a lo dicho es el caso de la discriminación por embarazo en el ámbito del empleo, discriminación relacionada con el permiso por maternidad o el sometimiento a un tratamiento de fecundación in vitro. Según la jurisprudencia consolidada del TJUE, cuando el perjuicio a la mujer es por estos motivos constituye una discriminación directa basada en el sexo y no resulta necesario identificar ninguna referencia comparativa. Vid., AGENCIA DE LOS DERECHOS FUNDAMENTALES DE LA UNIÓN EUROPEA (FRA) y CONSEJO DE EUROPA, *Manual de legislación europea contra la discriminación* (edición 2018), Luxemburgo, 2019.

²³⁶ *Manual de legislación europea contra la discriminación*, cit., p. 55.

²³⁷ Como se explica en el *Manual de legislación europea contra la discriminación*, cit., pp. 43 y ss, la normativa contra la discriminación afecta a nuestras decisiones de dos formas. En primer lugar, establece que

forma de discriminación se denomina «indirecta» porque no es el *trato* lo que difiere, sino sus *efectos*, que afectan de modo distinto a personas con características diferentes, situando a un grupo protegido en una desventaja particular. Es decir, la aplicación de la misma disposición, criterio o práctica a todas las personas sin tener en cuenta sus diferencias sustanciales es el motivo por el que se aprecia discriminación²³⁸. Por ello se viene manteniendo que la clave de la discriminación indirecta es que las situaciones diferentes deben recibir un trato diferenciado.

Se define este tipo de discriminación como aquélla que se produce cuando una disposición, criterio o práctica *aparentemente neutros*, ocasionen o puedan ocasionar una desventaja para una persona o grupo que comparten las mismas características (protegidas), sin que exista una justificación objetiva²³⁹.

las personas que se hallan en situación similar deben recibir un trato similar y no menos favorable por el simple hecho de tener una determinada característica «protegida» («discriminación directa»). En segundo lugar, en determinadas situaciones, el trato basado en criterios aparentemente neutros puede suponer un trato discriminatorio si resulta desfavorable para una persona o un grupo de personas en razón de una característica concreta de dicha persona o grupo («discriminación indirecta»).

²³⁸ Señala ARCOS VARGAS, M., *op.cit.*, p. 25, que estas situaciones pueden corregirse y evitarse mediante la adopción de medidas de acción positiva, es decir, medidas que tomando en consideración dichas diferencias sustanciales traten de lograr una igualdad real y no sólo formal. Similarmente, TOLOSA, P./DIBO, C., «Inteligencia artificial, discriminación por género y Derecho: viejos problemas, nuevos desafíos», en C. Danesi (dir.) *Inteligencia Artificial, tecnologías emergentes y Derecho*, 2021, p. 181. Según el TJUE, aunque dichas medidas puedan parecer discriminatorias, están diseñadas para eliminar o reducir las desigualdades de hecho que puedan existir en la sociedad.

²³⁹ Vid., por ejemplo, el art. 2.2.b) de la Directiva relativa a la igualdad racial y de la Directiva sobre igualdad en el empleo; el art. 2.1.b) de la Directiva sobre igualdad de trato entre hombres y mujeres (refundición) y el art. 2.b) de la Directiva sobre igualdad de trato entre hombres y mujeres en el acceso a los bienes y servicios. Para SÁEZ LARA, C., «El algoritmo como protagonista de la relación laboral. Un análisis desde la perspectiva de la prohibición de discriminación», *Temas laborales: Revista andaluza de trabajo y bienestar social*, N° 155, 2020, p. 48, el concepto de discriminación indirecta está dirigido a evitar el uso de criterios neutros pero con efectos discriminatorios, siempre que no estén justificados por su fin legítimo y no superen un escrutinio de adecuación y proporcionalidad. Y como afirma TOLOSA, P./DIBO, C., *op.cit.*, pp.180-181, suele distinguirse entre *trato desigual* e *impacto desigual* y los casos de impacto desigual se refieren a situaciones en las cuales se exigen ciertos requisitos que generan un efecto adverso hacia las personas que integran determinados grupos, aunque aparentemente no se trate de un trato desigual. Si el requisito que se exige, o la práctica en cuestión, tiene un efecto adverso hacia un determinado grupo, quien la requiere debe demostrar que la exigencia está justificada y es razonable pues tiene una relación estrecha con la función a desempeñar. Por su parte, el TC define la discriminación indirecta en su sentencia 61/2013, de 14 de marzo, como «aquel tratamiento formalmente neutro o no discriminatorio del que se deriva, por las diversas condiciones fácticas que se dan entre trabajadores de uno y otro sexo, un impacto adverso sobre los miembros de un determinado sexo» (por todas, STC 240/1999, de 20 de diciembre, FJ 6; o STC 3/2007, de 15 de enero, FJ 3), y recuerda en dicha resolución que tanto la jurisprudencia comunitaria como la suya propia, han destacado que «una norma puede dar lugar a discriminación indirecta “a menos que esté justificada objetivamente y sea proporcionada al objetivo perseguido”,

Se entenderá que existe dicha justificación cuando esa disposición, criterio o práctica persiga un fin legítimo y los medios para lograr ese fin sean apropiados y necesarios.

En definitiva, son dos los elementos que caracterizan la discriminación indirecta: en primer lugar, la adopción de una disposición, criterio o práctica aparentemente neutros; y en segundo lugar, que dicha adopción afecte de modo sustancialmente más perjudicial a un grupo protegido en comparación con otras personas en situación similar y que no pertenezcan a dicho grupo. Ambos elementos se recogen en el art.6.1 b) de la Ley 15/2022, que consagra el concepto de discriminación indirecta y la define como aquella que se produce «cuando una disposición, criterio o práctica aparentemente neutros ocasiona o puede ocasionar a una o varias personas una desventaja particular con respecto a otras por razón de las causas previstas en el apartado 1 del artículo 2».

Se trata de reaccionar frente al trato jurídico formalmente neutro, pero del que derivan consecuencias desiguales y perjudiciales por el *impacto* diferenciado y desfavorable que tiene sobre los miembros de uno y otro colectivo²⁴⁰. De ahí que a este tipo de discriminación se la conozca también como *discriminación de impacto* (*disparate impact*)²⁴¹. Igual que en los casos de discriminación directa, será preciso encontrar una referencia comparativa. Pero en estos casos, la comparación no opera entre individuos sino entre colectivos en los que se ponderan estadísticamente sus diversos componentes individuales.

considerando que “[a]demás, para estar justificada, la medida de que se trate tiene que ser adecuada para garantizar la realización del objetivo legítimo que persigue y no exceder de lo que es necesario para alcanzarlo” (STJUE de 13 de abril de 2010, *asunto Bressol y otros*)».

²⁴⁰ BARRERO ORTEGA, A., *op.cit.*, p. 49. En tal sentido, hay que valorar las consecuencias de la disposición o práctica y no sólo su neutra intención, de modo que si excluye o discrimina y no puede demostrarse su utilidad debe prohibirse. Cita este autor las siguientes sentencias del TJUE: 13 de mayo de 1986, *Bilka*; 17 de julio de 2014, *Leone y Leone*; y 18 de octubre de 2017, *Kalliri*.

²⁴¹ Así se denominan en la legislación norteamericana. La doctrina de nuestro país distingue entre la *discriminación directa* (o *intencional*) y la *indirecta* (o *por su efecto adverso*). En este sentido, entre otras, SÁEZ LARA, C., *op.cit.*, p. 48 o ALAMEDA CASTILLO, M.T., *op.cit.*, p. 13. No obstante, hay que aclarar que la discriminación indirecta puede ser tanto intencional como inconsciente o ignorada por quien discrimina. Para ALAMEDA CASTILLO, mientras en la directa hay voluntad de discriminar, en la indirecta no existe *—a priori—* intencionalidad discriminatoria, salvo que se conozca el impacto discriminatorio del algoritmo, se consienta y se siga usando. Entonces, la persona física o jurídica más próxima a la vulneración de derechos, al incumplimiento, será la que haya de asumir las consecuencias de la discriminación.

Sea como fuere, lo que parece claro es que la intención del discriminador es irrelevante para apreciar la existencia de discriminación. Lo decisivo es el efecto o impacto desproporcionado de la medida, práctica o disposición adoptada, esto es, su resultado discriminatorio. Sí será relevante esa intencionalidad a la hora de determinar la responsabilidad del autor, que lógicamente se verá agravada en ese caso. Así lo mantiene REY MARTÍNEZ, F., *Derecho Antidiscriminatorio*, Aranzadi, 2019, pp. 106-107.

En consecuencia, habrá que acreditar que un grupo resulta desfavorecido por una decisión respecto a otro grupo de referencia²⁴².

Si bien puede resultar relativamente sencillo identificar la discriminación directa, para probar la existencia de un caso *prima facie* de discriminación indirecta suelen necesitarse datos estadísticos en la medida en que se centra en el resultado y no tanto en el trato. Al analizar las pruebas estadísticas del efecto desproporcionadamente perjudicial para el grupo protegido respecto a otras personas en situación similar, el TJUE y el TEDH buscan muestras de que una proporción significativa de los perjudicados pertenezca a dicho grupo²⁴³.

Por lo que se refiere a la posibilidad de justificar la discriminación, si bien en el marco del CEDH la discriminación directa está sujeta a la causa de justificación general denominada *justificación objetiva*, en la legislación de la UE las causas de justificación frente a la misma son *limitadas*. Sin embargo, por lo que se refiere a la discriminación indirecta, todas sus formas están sujetas a la causa de justificación objetiva, con independencia de que la reclamación se

²⁴² *Manual de legislación europea contra la discriminación*, ap. 2.2.3, pp. 64 y ss. En la p. 64 se explica lo siguiente: «Tanto en los presuntos casos de discriminación directa como indirecta, los tribunales deben comparar, por ejemplo, los hombres con las mujeres, las parejas homosexuales con las parejas heterosexuales, o las personas con discapacidades con las personas sin discapacidades. Sin embargo, para la determinación de la discriminación indirecta, es necesario probar la existencia de dos grupos: uno en una posición de ventaja y otro en una posición de desventaja por la medida impugnada. En general, el grupo en situación de desventaja no está formado únicamente por personas con características protegidas. Por ejemplo, los trabajadores a tiempo parcial a los que coloca en una situación de desventaja determinada disposición son mayoritariamente mujeres, pero también puede haber hombres afectados. Por otra parte, no todas las personas con una característica particular sufren una situación de desventaja. Por ejemplo, en una situación en la que se exija un perfecto dominio de una lengua para poder acceder a un puesto, los candidatos extranjeros se encontrarán en general en una situación de desventaja, pero es posible que algunos de estos candidatos cumplan este requisito. En aquellos casos en los que un criterio formalmente neutro afecta en la práctica a un grupo completo, el TJUE ha apreciado que existe discriminación directa».

²⁴³ En el asunto *Di Trizio contra Suiza*, n. 7186/09, de 2 de febrero de 2016, el TEDH se basó en unas estadísticas que mostraban que el 97 % de las personas afectadas por el método de cálculo de la prestación por discapacidad eran mujeres que deseaban reducir su jornada laboral tras el nacimiento de un hijo. Vid., el *Manual de legislación europea contra la discriminación*, cit., p. 63, donde también se cita una recopilación de expresiones utilizadas por el TJUE que aparecen en las conclusiones del abogado general Léger en el asunto *Nolte* con respecto a la discriminación sexual (Conclusiones del abogado general Léger de 31 de mayo de 1995, apartados 57-58, en TJUE, C-317/93, *Inge Nolte/Landesversicherungsanstalt Hannover*, 14 de diciembre de 1995): «[P]ara presumirse discriminatoria, la medida debe afectar a un número mucho más elevado de mujeres que de hombres [Rinner-Kühn], o a un porcentaje considerablemente inferior de hombres que de mujeres [Nimz, Kowalska], o a un porcentaje muy superior de mujeres que de hombres [De Weerd]». Sobre los porcentajes exigidos para considerar probado el efecto desproporcionado de una disposición, práctica o medida, también puede verse SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación...», cit., p. 11/31.

realice al amparo del CEDH o de la legislación de la UE²⁴⁴. Ello implica que la discriminación indirecta sea más fácilmente justificable, ya que, mientras la directa sólo puede ser justificada en supuestos legalmente tasados, en la indirecta hay un mayor margen de justificación, por lo que, siempre que persiga una *finalidad legítima* y supere el *triple juicio de proporcionalidad* (esto es, idoneidad, necesidad y proporcionalidad en sentido estricto) podría justificarse²⁴⁵.

2.2.2. Discriminación por asociación, por error, múltiple e interseccional

Tanto los tribunales europeos como los nacionales han aceptado la llamada *discriminación por asociación* que tiene lugar cuando una persona recibe un trato menos favorable debido a su asociación o vinculación con otra que posee una característica protegida²⁴⁶. En estos casos la circunstancia personal o

²⁴⁴ ARCOS VARGAS, M., *op.cit.*, p. 22.

²⁴⁵ SORIANO ARNANZ, A., «La aplicación del marco jurídico europeo...», *cit.*, p. 77. Señala GARCÍA RUBIO, M.P., «Discriminación por razón de sexo y Derecho contractual en la Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva de mujeres y hombres», *Derecho Privado y Constitución*, n. 21, 2007 p.140, que el criterio de la *justificación objetiva* se concreta cumulativamente en que se persiga una «finalidad legítima» y que los medios para conseguirla sean «necesarios y adecuados», conceptos jurídicos todos ellos *a priori* indeterminados, pero que cuentan ya con una amplia tradición en las normas antidiscriminatorias y en la doctrina del TJUE. También explica NAVAS NAVARRO, S., «Negativa a contratar y prohibición de discriminar (Derecho comunitario y Derecho español)», ADC, tomo LX, 2007, fasc. IV, pp. 1630, que el principio de proporcionalidad se descompone a su vez en tres subprincipios: adecuación, necesidad y proporcionalidad en sentido estricto.

²⁴⁶ El TJUE reconoce la discriminación por asociación (tanto directa como indirecta) en los asuntos *C-303/06, S. Coleman v. Attridge Law y Steve Law [GS]*, 17 de julio de 2008 y *C-83/14, «CHEZ Razpredelenie Bulgaria» A. v. Komisia za zashtita ot diskriminatsia [GS]*, 16 de julio de 2015, respectivamente. En el caso Coleman, una mujer alegó un trato desfavorable en el trabajo debido a la discapacidad de su hijo. Esta discapacidad había provocado que, en ocasiones, llegara tarde al trabajo y que solicitara una reducción de la jornada laboral para atender a las necesidades de su hijo, solicitud que le había sido denegada. Además, había recibido amenazas de despido y comentarios despectivos respecto de la situación de su hijo. El TJUE consideró como referencia comparativa a sus compañeros de trabajo en puestos similares y con hijos, a los que sí se les ofrecía flexibilidad horaria cuando la solicitaban, por lo que determinó que el caso constituía un supuesto de discriminación y acoso por la discapacidad del hijo. En el caso Chez, un comerciante local que no pertenecía a la etnia gitana se vio perjudicado por el trato segregador contra la comunidad gitana de un barrio de Bulgaria en relación con la distribución de electricidad. El TJUE sostuvo que la discriminación por razón de la etnia se aplicaba a cualquier persona que, aun no perteneciendo a la raza o la etnia de que se tratara, sufriera un perjuicio por la medida discriminatoria en igual medida que las personas pertenecientes a la misma. Respecto a la referencia comparativa, el TJUE consideró que todos los usuarios finales de electricidad a los que un mismo distribuidor suministra, dentro de una entidad urbana con independencia del barrio en el que residan, se encuentran en una situación comparable. También el TEDH reconoció la discriminación por asociación en el asunto *Guberina contra Croacia* (n. 23682/13, 22 de marzo de 2016). Estimó que el trato discriminatorio del demandante debido a la discapacidad de su hijo era una forma de discriminación por motivo de discapacidad. En el caso concreto, las autoridades

social que motiva el trato peyorativo no concurre en la persona discriminada, sino en alguno de sus allegados o sujetos especialmente relacionados con ella (por ejemplo, se concede un trato peor a una trabajadora por la discapacidad de su hijo)²⁴⁷. Por ello, para apreciar esta discriminación no se ha de atender solo a un sujeto sino a dos, el perteneciente al colectivo vulnerable y el discriminado, de modo que se trata de una discriminación refleja²⁴⁸.

Distinta a la discriminación por asociación es la *discriminación por error* que se basa en una presunción sobre la persona que no es correcta (por ejemplo, se discrimina a una persona por su discapacidad o su homosexualidad cuando en realidad no concurre en ella dicha causa).

Ambos tipos de discriminación, que como se explicará después en el ámbito algorítmico pueden darse con frecuencia, han sido expresamente reconocidos en el art. 6.2 de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación (en adelante, Ley 15/2022 de igualdad). Su apartado a) define la *discriminación por asociación* como aquella que existe cuando una persona o grupo en que se integra, debido a su relación con otra en la que concurre alguna de las causas protegidas por la ley, es objeto de un trato discriminatorio; y su apartado b) dispone que la *discriminación por error* es aquella que se funda en una apreciación incorrecta acerca de las características de la persona o personas discriminadas.

Otros tipos de discriminación, que pueden ser directas o indirectas, son las discriminaciones *múltiples* y las *interseccionales*. La «discriminación múltiple» hace referencia a la discriminación que tiene lugar por diversos motivos protegidos que intervienen por separado (por ejemplo, se discrimina a una persona por su discapacidad y por su edad)²⁴⁹; mientras que la «discrimina-

des croatas denegaron una exención fiscal al demandante por la compra de un piso adaptado a las necesidades de su hijo con discapacidad (la familia tuvo que cambiar de casa porque necesitaba un piso con ascensor). El tribunal entendió que la administración croata no tuvo en cuenta el principio de acomodo razonable y accesibilidad de las personas con discapacidad. La jurisprudencia de distintos países también contempla este concepto. Vid., *Manual de legislación europea contra la discriminación*, cit., pp. 56 y ss. En España, la discriminación por asociación se reconoce expresamente en el art.2.e) y art. 63 del Real Decreto Legislativo 1/2013, de 29 de noviembre, por el que se aprueba el Texto Refundido de la Ley General de derechos de las personas con discapacidad y de su inclusión social y también en el art. 6.2.a) de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación.

²⁴⁷ Casos *Coleman v. Attridge Law* y *Steve Law y Guberina v. Croacia*, citados en la nota anterior.

²⁴⁸ BARRERO ORTEGA, A., *op.cit.*, p. 52.

²⁴⁹ La discriminación múltiple por dichos motivos se reconoce en la STC de 22 de enero de 2018, recurso de amparo 2699-2016, promovido por don A.R.S. en relación con las resoluciones de la Comunidad de Madrid y de las Salas de lo Contencioso-Administrativo del TSJ de Madrid y del TS que rechazaron su solicitud de reconocimiento de la situación de dependencia y determinación del programa individual de atención que le correspondía. El TC concede el amparo porque considera que concurre vulneración del derecho a no padecer discriminación por edad y discapacidad, ya

ción interseccional» se refiere a una situación en la que intervienen diversos motivos que interactúan entre sí de manera simultánea, de modo que resultan indisociables o inseparables (por ejemplo, ser menor de edad, mujer, pobre y tener VIH²⁵⁰) y *generan tipos específicos de discriminación*, lo que implica que dichos motivos no puedan analizarse por separado²⁵¹. Como apunta la doctrina, no cualquier combinación de factores permitirá identificar un grupo discriminado, sino que para apreciar discriminación interseccional deberá basarse en estereotipos sociales negativos que afectan a esa intersección, describiendo un resultado final, particular y propio, consecuencia de la sincrónica actuación de dos o más factores discriminatorios concurrentes²⁵².

que las resoluciones administrativas y judiciales aplicaron indebidamente una regla de exclusión por edad. En el caso concreto, no se asignó plaza al recurrente en una residencia de atención a personas con su tipo de discapacidad sólo porque era mayor de 60 años, asignándole plaza en una residencia ordinaria de personas mayores. Según el TC, hay discriminación por discapacidad en la medida en que se priva al recurrente de su derecho a la asistencia médica que precisaba su discapacidad, y también discriminación por edad porque se le impide el acceso al centro tan sólo por tener más de 60 años. En palabras de dicho Tribunal, la edad es «una segunda causa de discriminación que no desplaza, sino que se suma a la anterior –la discapacidad– (discriminación múltiple)».

²⁵⁰ Este fue el caso resuelto por la Corte interamericana de derechos humanos, en su sentencia González Lluy y otros v. Ecuador (2015) que resolvió el caso de una niña de 3 años contagiada de SIDA al recibir una transfusión sanguínea. Una de las consecuencias que se derivaron para ella fue la negativa a escolarizarla en un colegio. La Corte consideró que hubo discriminación interseccional por múltiples factores.

²⁵¹ *Manual de legislación europea contra la discriminación*, cit., p. 66. Un caso emblemático es el de una mujer negra a quien se le deniega la oportunidad de ser contratada como personal de fábrica en una empresa donde hay mujeres blancas trabajando en roles administrativos y hombres negros en puestos relacionados con la maquinaria, paquetería, etc. Si se analizara por separado el sexo, el género y la raza, podría parecer que la empresa no está discriminando por ninguna de dichas causas, ya que contrata tanto a mujeres como a personas negras. No obstante, la mujer negra se enfrenta a una situación única y distinta a la de las mujeres blancas y a la de los hombres negros, situación que precisamente es la base por la que no se la contrata. Así, por un lado, los estereotipos asociados al trabajo físico o manual le impiden acceder a puestos en la fábrica, mientras que, por otro lado, su falta de estudios la excluye de los roles administrativos. Vid., <https://wikigualdad.org/discriminacion-multiple>. Última consulta el 23 de abril de 2024.

²⁵² En este sentido, ABRIL STOFFELS, R., «El reconocimiento judicial en la discriminación múltiple en el ámbito europeo», *Revista de Derecho Comunitario Europeo*, n° 44, 2013, p. 321, que aporta como ejemplo de este tipo de discriminación el caso de *Beauty Salomon* (resuelto por la sentencia del TEDH de 25-7-2012, *B. S. v. España*, n. 47159/08) en el que ser mujer-prostituta-negra-extranjera genera una discriminación más profunda que no podría apreciarse por separado. En opinión de esta autora, es «[c]ierto que por ser mujer hay discriminación, por ejercer la prostitución, por ser de raza negra y por ser extranjera, cada uno de estos factores son determinantes de una vulnerabilidad hacia una discriminación, pero la vulnerabilidad de la persona y la posible discriminación resultante es mayor que la suma de las discriminaciones existentes. El elemento de la simultaneidad es clave, porque es lo que refuerza la discriminación y la convierte en algo más que una suma de discriminaciones». Sobre las discriminaciones múltiples e interseccionales, su diferencia, origen, etc., puede verse REY MARTÍNEZ, F., *op.cit.*, pp. 110-124 (y más concretamente, pp. 111-115). En la p. 114 se pone como

Estos tipos de discriminación también han sido recientemente reconocidos en la Ley 15/2022 integral de igualdad. En su art. 6.3 contempla la *discriminación múltiple* y la *interseccional* diferenciando ambas categorías de la siguiente manera: «a) Se produce discriminación múltiple cuando una persona es discriminada de manera simultánea o consecutiva por dos o más causas de las previstas en esta ley; b) Se produce discriminación interseccional cuando concurren o interactúan diversas causas de las previstas en esta ley, generando una forma específica de discriminación»²⁵³.

Tanto la discriminación por asociación o por error, como las discriminaciones múltiples o interseccionales, son particularmente relevantes en el ámbito algorítmico²⁵⁴. Piénsese que la ingente cantidad de datos procesados por estos sistemas y su gran precisión para detectar correlaciones entre ellos, podría llevar a que el algoritmo clasificara erróneamente a una persona, bien por su vinculación con otra, bien por equivocación; o podría ser que, debido a la granularidad con la que se tratan los datos, la discriminación llegara a producirse por la intersección de varias causas que por separado no la generarían. Volveremos sobre esta cuestión en el apartado 2.3.3.

Precisamente por ello, la consagración legislativa de estas modalidades discriminatorias supone un gran avance para afrontar en nuestro país los problemas que implica la discriminación algorítmica ante su escaso reconocimiento en el Derecho de la UE y, en consecuencia, por el TJUE²⁵⁵. En este

ejemplo de discriminación interseccional el de la mujer gitana. En este caso, la discriminación es el resultado de la combinación entre el género y la etnia, pero se trata de un resultado peculiar y genuino que debido a los particulares estereotipos sociales negativos es diferente de la discriminación que sufren las mujeres no gitanas y los hombres gitanos. También sobre la interseccionalidad en la discriminación puede verse el informe de Frías, M., «La interseccionalidad de la discriminación por razones de raza, etnia y género», financiado por la UE, Madrid, 2022. Recuperado de https://www.inclusion.gob.es/oberaxe/ficheros/documentos/I_Interseccionalidad_ES.pdf

²⁵³ También en el Preámbulo de esta Ley se hacen diversas referencias a estos tipos de discriminación poniendo énfasis en su gravedad y particular afectación a las mujeres a las que se considera «especialmente vulnerables a este tipo de discriminación», puesto que, dada la facilidad con la que se acumulan otras causas de discriminación a la discriminación contra la mujer, se generan nuevas formas de discriminación que conjugan el sexo o el género con la nacionalidad, la discapacidad, la edad, la religión, la orientación sexual o la condición socioeconómica. Así se explica en <https://wiki-igualdad.org/discriminacion-multiple>. Última consulta el 23 de abril de 2024.

²⁵⁴ Así lo afirman expertos como GERARDS, J./XENIDIS, R., *op.cit.*, pp. 63 o 65.

²⁵⁵ En el Derecho de la Unión, la única mención a la discriminación múltiple se encuentra en los considerandos de la Directiva sobre igualdad racial (2000/43/ CE) y de la Directiva relativa a la igualdad de trato en el empleo (2000/78/ CE), donde sucintamente se indica que «a menudo, las mujeres son víctimas de discriminaciones múltiples». Por su parte, el TJUE deniega la intersección en el contexto de la UE en el asunto *C-443/15, David L. Parris v. Trinity College Dublin y otros, de 24 de noviembre de 2016*. El TJUE considera que si una medida no constituye una discriminación por ninguno de los motivos prohibidos por la Directiva 2000/78/CE cuando dichos motivos se toman por sepa-

sentido, aunque el Derecho de la Unión admite que la discriminación puede basarse en varios motivos protegidos, el TJUE considera que no puede crearse ninguna categoría nueva de discriminación que sea el resultado de combinar varios de dichos motivos (por ejemplo, discriminación a mujeres que además sean negras), en la medida en que los motivos previstos en las Directivas de igualdad representan una lista exhaustiva²⁵⁶. A juicio de la doctrina, esta postura no es coherente con el impacto discriminatorio que tiene actualmente el uso de la IA en la toma de decisiones y en base a ella los autores consideran que el Derecho comunitario no está preparado para dar respuesta a la discriminación algorítmica en sus diversas manifestaciones debido a que adolece de vacíos legales que impiden cubrir las formas interseccionales de discriminación²⁵⁷.

Mayor reconocimiento ha tenido esta forma de discriminación por el TEDH, quien a pesar de no utilizar expresamente los términos «discriminación múltiple» ni «discriminación interseccional», tiene muy en cuenta este enfoque basado en múltiples motivos²⁵⁸. También en la Estrategia de Igualdad de Género 2020-2025, publicada por la Comisión Europea en marzo de 2020, se reconoce la interseccionalidad como un «principio transversal» e indica el propósito de que las políticas de la UE aborden la interseccionalidad del género con otros motivos de discriminación.

Por lo que se refiere al Derecho internacional, la interseccionalidad está reconocida oficialmente por el Comité de las Naciones Unidas para la Eliminación de la Discriminación contra la Mujer (Comité CEDAW) como un concepto necesario para comprender el alcance de la obligación de los Estados parte de erradicar la discriminación. El Comité indicó que «los Estados partes deben reconocer y prohibir en sus instrumentos jurídicos estas formas entrecruzadas de discriminación y su impacto negativo combinado en las mujeres afectadas»²⁵⁹.

rado, no puede dar lugar a una discriminación basada en el efecto combinado de dichos motivos; en este caso, la orientación sexual y la edad.

²⁵⁶ *Manual de legislación europea contra la discriminación*, cit., p. 70. Como se explica en esta obra, el TJUE ha reiterado en numerosas ocasiones que no le compete ampliar dichos motivos y, por ahora, no ha recurrido a los arts. 20 o 21 de la CDFUE para adoptar una postura diferente.

²⁵⁷ Así lo entiende RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», cit., p. 422.

²⁵⁸ Dos han sido los asuntos en los que el TEDH ha recurrido a la idea de la discriminación múltiple. En concreto, los resueltos por la sentencia de 25-7-2012, *B. S. contra España*, nº 47159/08, y por la sentencia de 25-7-2017, *Carvalho Pinto de Sousa Morais contra Portugal*, nº 17484/15.

²⁵⁹ Naciones Unidas (CEDAW), *Proyecto de Recomendación general nº 28 relativa al artículo 2 de la Convención sobre la eliminación de todas las formas de discriminación contra la mujer*, CEDAW/C/GC/28, 16 de diciembre de 2010, apartado 18: «El hecho de que los fenómenos de discriminación se superpongan es fundamental para el análisis del alcance de las obligaciones generales previstas en el artículo 2. La discriminación por razones de sexo o de género está inseparablemente ligada a otros factores

Pues bien, partiendo de que en el escenario algorítmico estas formas de discriminación adquieren especial relevancia debido a las características propias de estos sistemas, todos estos avances en el reconocimiento de las situaciones de discriminación basadas en la concurrencia de dos o más categorías protegidas, de la discriminación por asociación o de la discriminación por error, constituyen bases útiles para abordar algunos de los problemas que plantea a día de hoy la discriminación algorítmica²⁶⁰.

2.3. EL FENÓMENO DENOMINADO «DISCRIMINACIÓN ALGORÍTMICA»

2.3.1. Concepto de discriminación algorítmica

Llegados a este punto habría que concretar si el fenómeno sobre el que pivota el presente trabajo, esto es, la discriminación algorítmica, se presenta como una categoría jurídica nueva de discriminación, propia y distinta de la discriminación tradicional en la que no intervienen algoritmos, o si, por el contrario, no hay diferencias sustanciales entre ellas.

A tales efectos es preciso explicar previamente qué se viene entendiendo por discriminación algorítmica y cuáles son sus rasgos o características principales. La doctrina señala que es aquella que se produce cuando un individuo

como la raza, el origen étnico, la religión o las creencias, la salud, el estado civil, la edad, la clase, la casta y la orientación sexual y la identidad sexual. Puede afectar a las mujeres pertenecientes a estos grupos en diferente grado o de forma distinta que a los hombres. Los Estados partes deben prevenir legalmente estas formas superpuestas de discriminación y el efecto acumulativo de sus consecuencias negativas para las personas, y deben prohibirlas. Asimismo, deberán adoptar e implementar políticas y programas destinados a eliminar estas formas de discriminación, y tomar medidas especiales de carácter temporal cuando sea necesario de conformidad con el artículo 4, párrafo 1, del Convenio y la recomendación general N.º 2».

²⁶⁰ GERARDS, J./XENIDIS, R., *op.cit.* p. 142. Para estos autores, a pesar de que el TJUE consideró que no existía discriminación interseccional en el caso *David L. Parris v. Trinity College Dublin y otros*, anteriormente citado, ha sugerido que reconocer el concepto de interseccionalidad podría enriquecer las evaluaciones de discriminación e incluso ha reconocido esta modalidad de discriminación de forma implícita en alguna ocasión. Por ejemplo, en sus sentencias de 6 de diciembre de 2012, *Johann Odar v Baxter Deutschland GmbH*, C-152/11, EU:C:2012:772 y 19 de septiembre de 2018, *Surjit Singh Bedi v Bundesrepublik Deutschland and Bundesrepublik Deutschland in Prozessstandschaft für das Vereinigte Königreich von Großbritannien und Nordirland*, C-312/17, EU:C:2018:734, reconoce que la interacción de la discriminación basada en la edad y la discapacidad genera una especial desventaja y el riesgo de que las personas gravemente discapacitadas puedan tener necesidades financieras como consecuencia de su discapacidad que no pueden ajustarse y que, a medida que aumente su edad, se verán agravadas. Sobre el tema puede verse también XENIDIS, R., «Tuning EU equality law to algorithmic discrimination: Three pathways to resilience», *Maastricht Journal of European and Comparative Law*, n. 27 (6), 2020, pp. 736-758. <https://doi.org/10.1177/1023263X20982173>, p. 744.

o grupo recibe un tratamiento arbitrario o injusto como consecuencia de la toma de decisiones automatizadas²⁶¹. A pesar de su concisión (o más bien en base a ella) podría concluirse que la única diferencia con la discriminación que podríamos llamar «analógica» sería la forma de llevarla a cabo, en tanto que aquella implica el uso de algoritmos y también, usualmente, de tecnología inteligente. Pero la conducta discriminatoria sería idéntica en ambos casos siempre que la decisión adoptada (bien por el humano, bien por el sistema de IA) diera lugar a un trato diferente e injusto en situaciones comparables, o a un trato idéntico en situaciones distintas, basado en motivos protegidos y sin que hubiese justificación objetiva y razonable para ello.

Ahora bien, a nuestro modo de ver, esto no significa que la vía (analógica o digital) a través de la que se acaba discriminando sea una cuestión menor. Muy al contrario, el empleo de sistemas de IA para la toma de decisiones implica que la discriminación adquiera unas dimensiones y afronte unos retos hasta el momento inexistentes²⁶². Basta pensar que, si un sistema algorítmico es capaz de almacenar, procesar y analizar cantidades de datos inasumibles para el cerebro humano a una velocidad sin precedentes, el número de decisiones potencialmente discriminatorias que podría arrojar dicho sistema

²⁶¹ Así, MERCADER UGUINA, J.R. «Discriminación algorítmica y derecho granular: nuevos retos para la igualdad en la era del big data», *Labos: Revista de Derecho del Trabajo y Protección Social*, vol. 2, n. 2, 2021, p. 6. Con más detalle la define CASTILLO OCARANZA, C., «Discriminación algorítmica en el ámbito laboral», en P. Rivas Vallejo (dir.), *Discriminación algorítmica en el ámbito laboral*, Navarra 2022, p. 72, para quien existe discriminación algorítmica cuando un «algoritmo o sistema computacional algorítmico (X) produce discriminación grupal para una persona (Y) respecto de otra (Z), es decir, produce un resultado peor para la persona (Y) respecto a la persona (Z) debido a que (Y) pertenece a un cierto grupo social saliente o «socially salient group» (entendiendo por tal, aquél que pueda ser fácilmente identificado y que sea tal que pertenecer o no a este grupo tenga consecuencias para las relaciones sociales, y en particular, las relaciones de poder –ej. grupos delimitados por género, etnia, raza, discapacidad etc). Además, (X) debe haber basado su decisión en una observación estadísticamente relevante, es decir, en cualquier información que (X) utilice basándose en los datos de entrada (ej. si un empleador no contrata a una persona por tener la característica de ser mujer –que es un grupo social saliente– y contrata a otra que no es mujer basándose en la observación estadística de que es más probable que una mujer pida una baja por maternidad»; FERNÁNDEZ DE LA MORENA, B., *op.cit.*, p.1, señala que «La discriminación algorítmica tiene lugar cuando un individuo o colectivo recibe un tratamiento injusto como consecuencia de la toma de decisiones algorítmica automatizada»; y CASTELLANOS CLARAMUNT, J., «Derecho e inteligencia artificial: atención especial a los sesgos, la privacidad y la protección de datos», accesible en <https://idpbarcelona.net/derecho-e-inteligencia-artificial-atencion-especial-a-los-sesgos-la-privacidad-y-la-proteccion-de-datos/>, la define como «un fenómeno que se produce cuando sistemas de IA o algoritmos utilizados para tomar decisiones automatizadas discriminan de manera indirecta o inconsciente a ciertas personas o grupos».

²⁶² En este sentido, por todos, GINÉS i FABRELLAS, A., *op.cit.*, pp. 316 y ss. También GERARDS, J./ XENIDIS, R., *op.cit.*, p. 75, ponen de manifiesto que, si bien muchos de los problemas que plantea la discriminación existían antes del desarrollo de los algoritmos y han sido identificados por la doctrina, el uso creciente de algoritmos en todos los ámbitos de la sociedad los agudiza y multiplica.

aumentaría exponencialmente respecto a las que podría adoptar una persona no provista de tales herramientas²⁶³. Además, la granularidad de los datos permitiría al algoritmo encontrar correlaciones e inferencias entre la información suministrada, difícilmente detectables para las personas, e incluso en ocasiones, inesperadas o irracionales para ellas.

Esto ha dado lugar al fenómeno conocido como la *automatización de la discriminación*, o la capacidad de estos sistemas para convertir la discriminación en constante, uniforme e inflexible, reproduciéndola a una escala y velocidad vertiginosas²⁶⁴. Así, la doctrina viene advirtiendo reiteradamente desde hace años sobre la potencialidad de la tecnología para *extraer y procesar datos, reproducir y perpetuar discriminaciones existentes, amplificar sesgos y hacer que la discriminación pase desapercibida*²⁶⁵.

De ahí que se hayan destacado tres rasgos que caracterizan a la discriminación algorítmica: su invisibilidad, su intensidad y su complejidad técnica²⁶⁶.

²⁶³ Señala CASTILLO OCARANZA, C., *op.cit.*, p. 75, en base al trabajo de O'Neil, que uno de los tres principales riesgos de la discriminación algorítmica es su «escala». Piénsese, como explica el autor, en que el daño que puede hacer un empleador racista está acotado a las oportunidades de decisión que tenga, las cuales, a su vez, están limitadas por la cantidad de decisiones de contratación, promoción, etc, que realiza en un cierto periodo de tiempo. Sin embargo, los sistemas computacionales se utilizan normalmente para procesar un gran volumen de datos, por lo que, en el caso de que sean discriminatorios, puede producir un daño a una escala mucho mayor, proporcional al número de casos que procesa. En sentido similar, ORWART, C., *op.cit.*, p. 22, al comparar las decisiones humanas con las de los sistemas automatizados, hace hincapié en el número de decisiones que pueden adoptarse por una y otra vía, concluyendo que el riesgo de discriminación a través de estos sistemas puede convertirse en un fenómeno de masas y conducir fácilmente a desventajas acumulativas.

²⁶⁴ EUBANKS, V., *op.cit.* y O'NEIL, C., *op.cit.* Esta última autora cuenta que si un estudiante pobre no puede obtener un préstamo porque un modelo de IA, en virtud de su código postal, lo considera demasiado arriesgado, se le estaría privando de un tipo de educación que podría sacarlo de la pobreza, con lo que se acaba produciendo una espiral viciosa, maquiavélica y discriminatoria que impide que toda una clase social desfavorecida pueda salir del hoyo. Y entre nuestros autores, explica por todos SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación...», cit., p. 11/31, que cuando un algoritmo tiene un sesgo que perjudica a personas o grupos protegidos, a menos que el sistema incluya directrices para corregir estos sesgos, lo normal será que dicho sesgo persista durante todo el tiempo en que el sistema esté en funcionamiento.

²⁶⁵ Peligros todo ellos explicados con detalle por GINÈS i FABRELLAS, A., *op.cit.*, pp. 315-321 y doctrina citada por la autora entre la que se encuentran los trabajos de BOLUKBASI y otros, o ZHAO y otros, ya citados en esta obra. También RAMÍREZ-BUSTAMANTE, N./ PÁEZ, A., *op.cit.*, pp. 9 y ss; BAROCAS, S./SELBST, A. D., *op.cit.*, p. 674. En la *Recomendación sobre la ética de la IA* de la UNESCO también se presta especial atención a los retos que plantean las decisiones algorítmicas porque pueden «reproducir y reforzar los sesgos existentes, lo que puede exacerbar las formas ya existentes de discriminación, los prejuicios y los estereotipos». Y en la misma línea se posiciona el Informe *Algorithms and Human Rights*, elaborado por el Comité de Expertos en intermediarios de internet del Consejo de Europa (MSI-NET), pp. 27-28.

²⁶⁶ Rasgos identificados por SÁEZ LARA, C., *op.cit.*, p. 45 y ALAMEDA CASTILLO, M.T., *op.cit.* p. 14, siguiendo a la primera. Según opina ESPÍN ALBA, I., *op.cit.*, p. 137, el sesgo algorítmico es espe-

Según mantiene un sector doctrinal al que nos adherimos, dichas características y sus consiguientes desafíos, justifican que la discriminación algorítmica sea un fenómeno nuevo y distinto que obliga a llevar a cabo dos cometidos:

Por una parte, a hacer *ajustes en los mecanismos de tutela antidiscriminatoria* existentes incorporando exigencias de transparencia y explicabilidad desde el diseño, evaluaciones de impacto que midan el potencial de discriminación del sistema por diversas causas (género, etnia, religión etc.) y auditorías, con el fin de poder detectar la discriminación, prevenirla o corregirla y, en última instancia, reparar los daños que ocasione²⁶⁷. No porque se trate de una categoría jurídica de discriminación propia y distinta de las anteriormente explicadas, sino porque las novedades tecnológicas que presenta (opacidad, complejidad, dependencia de datos, etc.) hacen que alcance unas dimensiones y una problemática no previstas hasta ahora²⁶⁸.

Por otra parte, a actualizar y ampliar el ámbito de aplicación objetivo y subjetivo del Derecho comunitario antidiscriminatorio porque es incompleto y fragmentario por varias razones: en primer lugar, porque actualmente el uso de estos sistemas algorítmicos no solo se lleva a cabo en los sectores que cubren las denominadas Directivas de igualdad (por ejemplo, el acceso a bienes

cialmente peligroso para los derechos fundamentales por el volumen de datos manejados y el impacto –en muchas ocasiones invisible– en las más diversas áreas de actuación de las personas, algunas de ellas especialmente sensibles. También en esta línea GERARDS, J./ XENIDIS, R., *op.cit.*, p. 108.

²⁶⁷ Así lo venía manteniendo la Agencia Europea de los Derechos Fundamentales tal como pone de manifiesto SÁEZ LARA, C., *op.cit.*, p. 58. Apunta la autora, que esa transparencia y evaluación debe verificarse tanto en relación a los datos de los que se nutre el sistema (los cuales han de ser de calidad), como sobre la forma en que funcionan los algoritmos (caja negra) y sus facultades predictivas. Respecto a la calidad de los datos, es esencial recopilar metadatos (es decir, información sobre los datos) y evaluar su corrección y generalización, ya que esto puede ayudar a comprender y mitigar los efectos discriminatorios de estos sistemas. A mayor abundamiento, esa evaluación de los algoritmos debe extenderse también al desarrollo de los algoritmos y a la forma en que trabajan. Por tanto, la calidad de los datos junto con el diseño de los algoritmos y los procesos de reevaluación constante, deberían evitar la aparición de sesgos. Para MERCADER UGUINA, J.R., *op.cit.*, p. 6, aunque por el momento la discriminación algorítmica obtiene respuestas desde nuestras actuales técnicas conceptuales, requiere que se adopten nuevas herramientas jurídicas de análisis. Y en opinión de HACKER, P. *op.cit.*, p. 27, teniendo en cuenta que los derechos de acceso individual a los datos establecidos en el RGPD no son suficientes para garantizar la equidad, las auditorías y evaluaciones de impacto se tornan necesarias para abordar los sesgos estructurales en los sistemas algorítmicos. Como se explicará después, algunas de estas medidas las ha recogido el reciente RIA, en particular, las dirigidas a garantizar la transparencia de los sistemas y su auditoría. La Ley 15/2022, integral de igualdad, también recoge la obligación de adoptar estos mecanismos en el ámbito de la IA.

²⁶⁸ Para GINÉS i FABRELLAS, A., *op.cit.*, p. 312, no es necesario crear categorías jurídicas nuevas de discriminación puesto que, sin perjuicio de las novedades técnicas que introduce ahora la toma de decisiones a través de sistemas de IA, es posible aplicar la actual doctrina antidiscriminatoria a la discriminación algorítmica.

y servicios o las relaciones laborales), sino en todos los ámbitos y actividades de la sociedad; en segundo lugar, por la potencialidad de la discriminación algorítmica para dar lugar a nuevas formas de discriminación basadas en motivos no reconocidos expresamente como «protegidos» en el Derecho europeo como la renta o la clase social; y en tercer lugar, porque generaliza formas de discriminación como la discriminación por asociación y por error, e incluso algunos tipos escasamente reconocidos en la legislación comunitaria como la discriminación múltiple e interseccional²⁶⁹.

Esta situación en que se encuentra el Derecho comunitario hace que sea necesario y prioritario que legisladores y aplicadores del Derecho le presten una mayor atención. En este sentido, es importante señalar que en nuestro ordenamiento algunos de estos inconvenientes han sido soslayados con la entrada en vigor de la Ley 15/2022 integral de igualdad, que ha ampliado considerablemente el elenco de características protegidas por las Directivas europeas y junto a los seis motivos reconocidos en ellas (sexo, origen racial o étnico, discapacidad, edad, religión o creencias y orientación sexual), y tomando como referencia el art. 14 CE, incorpora expresamente los de *enfermedad o condición de salud, estado serológico y/o predisposición genética a sufrir patologías y trastornos, identidad sexual, expresión de género, lengua y situación socioeconómica*, y mantiene la cláusula de cierre del citado precepto constitucional («*cualquier otra condición o circunstancia personal o social*»); además, por lo que se refiere al ámbito de aplicación objetivo, el carácter integral de esta Ley se manifiesta también en los ámbitos de la vida política, económica, cultural y social a los que se aplica: a saber, *el empleo, el trabajo, la educación, la sanidad, servicios sociales, el acceso a bienes y servicios, incluida la vivienda, la participación social o política y la publicidad y medios de comunicación*, estableciendo un conjunto de obligaciones que vinculan incondicionadamente a todas las administraciones públicas y en la forma que la propia Ley establece en el caso de las relaciones entre particulares²⁷⁰.

Por último, reconoce expresamente los tipos de discriminación más relevantes que pueden ocasionarse a través del uso de sistemas de IA para la toma de decisiones, como la discriminación múltiple, la interseccional, por asociación y por error. Con lo cual, puede afirmarse que esta Ley consigue cubrir, al menos en nuestro territorio, algunas de las lagunas de las que adolece el Derecho comunitario para hacer frente a la discriminación algorítmica²⁷¹.

²⁶⁹ Algunas de las razones que se esgrimen para afirmar el incremento de estos tipos de discriminación en el ámbito algorítmico son, por ejemplo, la ingente cantidad de datos analizados y procesados por estos sistemas, la granularidad de tales datos, o los *proxies* o correlaciones entre puntos de datos. Vid., GERARDS, J./ XENIDIS, R., *op.cit.*, pp. 111 y ss.

²⁷⁰ Así se recoge en su Exposición de Motivos.

²⁷¹ Entre los propósitos de la Ley está el de transponer adecuadamente los objetivos y fines de las Directivas de Igualdad, lo que solo se hizo parcialmente con la Ley 62/2003 de medidas fiscales, administrativas y del orden social, de manera insuficiente e ineficiente.

2.3.2. Discriminación algorítmica y sesgo algorítmico

Es preciso aclarar que «discriminación» y «sesgo» no son conceptos idénticos o coincidentes a pesar de que con frecuencia se empleen indistintamente para hacer referencia al fenómeno que venimos analizando.

El término «sesgo» tiene diversos significados y se utiliza con distintos alcances. En un primer sentido, vendría a ser sinónimo de *prejuicio*. En otras ocasiones, adquiere un sentido equivalente al de *estereotipo*. Además, cuando decimos que una opinión o conducta adolece de sesgos, nos referimos a que la misma está *influenciada* o *manipulada*²⁷². Como puede observarse con facilidad, en todos estos casos la palabra «sesgo» generalmente implica una connotación negativa. Y es precisamente este sentido el que suele darse al sesgo cuando se utiliza en el ámbito jurídico, esto es, como sinónimo de *prejuicio* o *trato injusto*. Dicha connotación es la que se le otorga en la Propuesta de Reglamento sobre aspectos éticos de la IA, cuyo art. 4.1) define el «sesgo» como «toda percepción personal o social prejuiciosa de una persona o de un grupo de personas sobre la base de sus características personales»²⁷³.

Ahora bien, cuando se habla concretamente de *sesgo algorítmico*, *sesgo del algoritmo* o *sesgo en los modelos de machine learning*, su significación es mucho más amplia e implica cualquier desviación inadecuada en un proceso de inferencia²⁷⁴, no necesariamente discriminatoria en sentido jurídico. En el ámbito de la *estadística* suele emplearse para describir un factor que cambia un resultado de cierta forma, o también para hacer referencia a aquellos inconscientes emanados de las personas involucradas en el diseño de los sistemas o implícitos en los datos²⁷⁵.

Como explican GERARDS y XENIDIS, históricamente, los conceptos de «sesgo» (*bias* en inglés) y «justicia/equidad» (*fairness*) han sido utilizados en

²⁷² TOLOSA, P./DIBO, C., *op.cit.*, p. 168.

²⁷³ Dicha Propuesta acompaña a la *Resolución del Parlamento Europeo, de 20 de octubre de 2020, con recomendaciones destinadas a la Comisión sobre un marco de los aspectos éticos de la inteligencia artificial, la robótica y las tecnologías conexas* (2020/2012(INL)) y define por un lado el *sesgo* en el art. 4, apartado l) y la *discriminación* en el apartado m) como «todo trato diferenciado de una persona o de un grupo de personas basado en un motivo que no tiene justificación objetiva o razonable alguna y que, por tanto, está prohibido por el Derecho de la Unión».

²⁷⁴ AEPD, *Informe sobre la Adecuación del RGPD a la IA*, cit., p. 7.

²⁷⁵ MUÑOZ GUTIÉRREZ, C., *op.cit.*, p. 279. Según dispone la séptima acepción de la RAE: «*Estad.* Error sistemático en el que se puede incurrir cuando al hacer muestreos o ensayos se seleccionan o favorecen unas respuestas frente a otras». Señala FERNÁNDEZ DE LA MORENA, B., *op.cit.*, p. 2, que conceptos como el sesgo y la discriminación, a pesar de no ser conceptos análogos, están muy relacionados y que el sesgo se encuentra tradicionalmente asociado a una representación desigual de las clases en una base de datos.

el ámbito tecnológico o informático para describir una serie de problemas éticos relacionados con el funcionamiento y resultado de los algoritmos. Específicamente, el «sesgo algorítmico» se refiere a «un error sistemático» *de cualquier tipo* en el resultado de las operaciones o decisiones algorítmicas. Por lo tanto, el concepto de «sesgo» tiene un significado mucho más amplio que el concepto de «discriminación» que utilizamos en el ámbito jurídico, ya que hace referencia a todo tipo de errores sistemáticos, ya sean errores estadísticos, cognitivos, sociales, estructurales o institucionales. Sin embargo, cuando se emplea este término en el contexto de la justicia o la equidad, el «sesgo algorítmico» se refiere a un tipo particular de error que sitúa a los grupos privilegiados en una posición de ventaja y a los no privilegiados en una posición de desventaja sistemática. Aquí se aproxima en mayor medida al concepto de discriminación que empleamos en el ámbito jurídico, entendida como el trato injusto o desfavorable que se otorga a una persona o grupo en relación a otro en base a ciertos motivos, o el impacto desproporcionadamente desventajoso que una medida o política adoptada puede tener en un grupo específico²⁷⁶.

Aun siendo esto así, añaden estos autores que en cualquier caso el término «sesgo algorítmico» sigue siendo más amplio que el de «discriminación algorítmica», por cuanto que ésta se aplica exclusivamente al trato desfavorable injustificado o desventaja experimentada por categorías específicas de población protegidas por la ley, bien de forma explícita (por ejemplo en base a ciertos motivos protegidos) o bien implícitamente (por ejemplo mediante cláusulas de no discriminación generales o abiertas), mientras que aquél se refiere a cualquier tipo de desventaja que podría considerarse ética o moralmente incorrecta²⁷⁷.

Los sesgos, conscientes o inconscientes, son particularmente graves cuando derivan en discriminaciones de personas o grupos no toleradas por la ley. En tales sesgos centraremos nuestra atención. Por tanto, a los efectos del presente trabajo, no interesa cualquier tipo de desviación inadecuada del pro-

²⁷⁶ Vid., GERARDS, J./XENIDIS, R., *op.cit.*, p. 49.

²⁷⁷ Por ejemplo, un algoritmo que pone en desventaja a los grupos de bajos ingresos y privilegia a las personas con altos ingresos podría verse como una forma de sesgo algorítmico desde un punto de vista ético, pero no se contempla como característica protegida en las Directivas europeas. No obstante, para ESPÍN ALBA, *op.cit.*, p. 146, el tratamiento de la discriminación en el ámbito algorítmico abarcaría tanto la discriminación directa e indirecta establecida en el marco legal antidiscriminatorio, que hace referencia al tratamiento diferente y peor de *ciertos colectivos* respecto de otros (p.ej. no se concede un crédito a una mujer, mientras se concede a un hombre con situación de solvencia similar), como otras situaciones en las que, debido a las correlaciones entre los datos y parámetros utilizados por los algoritmos para tomar decisiones, se brinda un trato distinto a una persona respecto de otra u otras, *sin que necesariamente pertenezca o esté vinculada con un grupo vulnerable* (p.ej., ser categorizado como un *heavy spenders* por un historial de compras que sugiere ese patrón de comportamiento o por el acceso a las compras en línea con dispositivos de una determinada marca).

ceso de inferencia (sesgo algorítmico en general), sino sólo aquellos sesgos problemáticos desde el punto de vista del derecho a la igualdad y no discriminación (sesgos discriminatorios en particular). Es decir, interesan aquellos que conllevan prejuicios o estereotipos que dan lugar a decisiones automatizadas que perjudican directa o indirectamente a una persona o grupo respecto de otro, generando discriminación prohibida.

Huelga mencionar que el problema de los sesgos no es exclusivo de las decisiones algorítmicas, sino de cualquier proceso de toma de decisión, ya sea humano o automatizado²⁷⁸. En ambos casos los sesgos son igualmente relevantes, pero como apunta ATIENZA NAVARRO parece que las personas toleramos peor las vulneraciones de derechos ocasionadas por sistemas de IA que por seres humanos²⁷⁹. Además, ambos pueden ser oscuros. También la mente humana puede ser una caja negra que impida llegar a saber los verdaderos motivos por los cuales se adoptan decisiones. Es decir, puede ser imposible saber por qué realmente un juez decidió imponer la máxima pena permitida a un acusado, por qué un empleador rechazó a un candidato a un puesto de trabajo, o por qué un arrendador no alquiló su piso a quien se interesó por él. Podríamos hablar en estos casos de «discriminación oculta», es decir, de aquella que acaece cuando el discriminador encubre los motivos reales del trato distinto y peor que otorga a una persona o grupo, y que se basa en prejuicios machistas, homófobos, racistas, etc., alegando una causa de diferenciación en principio válida e ideológicamente neutra, ajena a tales motivos²⁸⁰.

En la literatura se han identificado numerosos tipos de sesgos (estadísticos, cognitivos, culturales, etc.)²⁸¹. Uno especialmente preocupante en este ámbito es el denominado «sesgo de la automatización» que consiste en que el humano interprete o asuma acríticamente los resultados de un sistema inteligente como verdades incontestables y absolutas, aceptando un principio de

²⁷⁸ AEPD, *Informe sobre la Adecuación del RGPD a la IA*, cit., pp. 7 y p. 37. Para TOLOSA, P./DIBO, C., *op.cit.*, p. 163, incluso quienes tengan las mejores intenciones pueden tener sesgos inconscientes e implícitos contra cierto grupo. En p. 189 señalan que casos como el de Amazon generan escándalo y ponen el foco en los riesgos que la inteligencia artificial puede tener a la hora de exacerbar la discriminación en razón del género. Sin embargo, esa clase de problemas también se dan cuando el proceso de decisión no involucra, en ninguna instancia, algoritmos.

²⁷⁹ ATIENZA NAVARRO, M.L., «La responsabilidad civil por daños causados por inteligencia artificial. Estado de la cuestión», en *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, coord. por N. Álvarez Lata; J.M. Busto Lago (pr.), 2024, p. 346.

²⁸⁰ REY MARTÍNEZ, F., *op.cit.*, p. 106. Para TOLOSA, P./DIBO, C., *op.cit.*, p. 159, normalmente el motivo subyacente a la diferencia de trato está en la mente de su autor y la información y los archivos que podrían servir para demostrarla suelen estar en manos de la persona a la que se dirige el reproche de discriminación.

²⁸¹ Puede verse la clasificación llevada a cabo por ESPIN ALBA, I., *op.cit.*, pp. 137-148, que recoge los sesgos más habituales y la doctrina citada por dicha autora.

autoridad derivado de las expectativas creadas por dichos sistemas²⁸². Tanto si el humano adopta una decisión basada exclusivamente en el resultado discriminatorio que arroja el sistema, como si es el propio sistema el que decide y ejecuta automáticamente la decisión discriminatoria, habría discriminación. Por ello, como se dijo, tanto en los sistemas de decisión *plenamente automatizados* como en los *semiautomatizados o de apoyo a la decisión* existe un riesgo de vulnerar el derecho a la igualdad y no discriminación. Así lo reconoce el Parlamento Europeo en su *Propuesta de Reglamento sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas*, al señalar que «El sesgo puede originarse tanto en las decisiones basadas en un sistema automatizado o adoptadas por tal sistema, como en los conjuntos de datos en los que se basan dichas tomas de decisiones o con los que se entrena al sistema»²⁸³, haciendo referencia con ello a algunas de las causas o factores por las que el algoritmo puede llegar a resultados discriminatorios²⁸⁴.

Nótese que de todas las Resoluciones y Propuestas europeas en materia de IA, sin duda es ésta la que presta mayor atención al problema de los sesgos discriminatorios. Prueba de ello es que ambos conceptos («sesgo» y «discriminación») aparecen de forma inseparable y continua a lo largo del texto. Así, por ejemplo, cuando el Parlamento declara la *ilicitud* del «sesgo y la discriminación de los programas informáticos, los algoritmos y los datos», advirtiendo sobre la necesidad de abordar el problema a través de la regulación de los procesos mediante los cuales se diseñan y despliegan.

También se explica en dicha Resolución cuándo se consideran sesgados estos programas, algoritmos o datos y cuándo han de considerarse discriminatorios. Se consideran sesgados, por ejemplo, cuando *no arrojen resultados óptimos en relación con cualquier persona o grupo de personas, sobre la base de una percepción personal o social prejuiciosa y del posterior tratamiento de los datos relativos a sus características*. Y deben considerarse discriminatorios, en consonancia con el Derecho de la Unión, *cuando produzcan resultados con unos efectos negativos desproporcionados que den lugar a un tratamiento diferente de una persona o grupo de personas, que las ponga incluso en una situación de desventaja en comparación con otras personas, sobre la base de motivos tales como sus características personales, sin una*

²⁸² Destaca este sesgo como especialmente peligroso, la AEPD en su *Informe sobre la Adecuación del RGPD a la IA*, cit., p. 7. También se recoge expresamente en el RIA (art. 14.4.b) donde se dice que las personas físicas que vigilen el sistema han de «ser conscientes de la posible tendencia a confiar automáticamente o en exceso en los resultados de salida generados por un sistema de IA de alto riesgo («sesgo de automatización»), en particular con aquellos sistemas que se utilizan para aportar información o recomendaciones con el fin de que personas físicas adopten una decisión».

²⁸³ Considerando 22.

²⁸⁴ Estas causas serán analizadas en el siguiente apartado.

*justificación objetiva o razonable e independientemente de cualquier reivindicación de neutralidad de las tecnologías en cuestión*²⁸⁵.

A fin de evitarlos, establece su art. 9 la obligación de que no exista sesgo ni discriminación, disponiendo de forma imperativa que dichos programas, algoritmos o datos *serán imparciales y no discriminarán* por motivos de raza, sexo, orientación sexual, embarazo, discapacidad, características físicas o genéticas, edad, minoría nacional, origen étnico o social, lengua, religión o creencias, opiniones políticas o participación cívica, nacionalidad, estado civil o económico, educación o antecedentes penales²⁸⁶, ampliando considerablemente el listado de características protegidas por las Directivas europeas.

2.3.3. La discriminación algorítmica como discriminación indirecta

Es opinión ampliamente compartida entre los expertos que la mayoría de casos de discriminación algorítmica han de calificarse como discriminaciones *indirectas*, sin perjuicio de que también puedan darse de forma excepcional supuestos de discriminación algorítmica *directa*²⁸⁷. El principal in-

²⁸⁵ Considerandos 23 y 24.

²⁸⁶ Añade su segundo apartado que «solo podrá justificarse el trato diferenciado entre personas o grupos de personas cuando exista una finalidad objetiva, razonable y legítima que sea proporcionada y necesaria en la medida en que no exista otra alternativa que interfiera menos con el principio de igualdad de trato». Además, como objetivos o finalidades legítimas para justificar objetivamente cualquier diferencia de trato, señala los siguientes: la protección de la seguridad y la salud públicas, la prevención de infracción penales, la protección de los derechos y las libertades fundamentales, la representación equitativa y los requisitos objetivos para ejercer una profesión.

²⁸⁷ Entre otros muchos, GINÉS I FABRELLAS, A., *op.cit.*, pp. 312-3; CASTILLO OCARANZA, C., *op.cit.*, p. 73; SÁEZ LARA, C., *op.cit.*, p. 48; MUÑOZ GUTIÉRREZ, C., *op.cit.*, p. 285. También en el estudio de GERARDS, J./XENIDIS, R., *op.cit.*, principalmente en el capítulo 2 y sus conclusiones, se ponen de relieve varias cuestiones sobre esta idea: la primera, que en la literatura se considera poco probable que la discriminación algorítmica sea directa porque esto implicaría introducir expresamente en los modelos los motivos protegidos como variables negativas, cosa que no suele ocurrir; en segundo lugar, que los expertos afirman que los sesgos en el procesamiento de datos dan lugar a formas indirectas de discriminación en lugar de directas y por ello hay consenso académico en que el concepto de discriminación indirecta encajaría mejor con los modelos algorítmicos que la discriminación directa; en tercer lugar, que en su opinión ambos argumentos deben matizarse en la medida en que la discriminación algorítmica desdibuja los límites entre la discriminación directa e indirecta, lo que lleva a un problema a la hora de calificar el caso concreto y, dado que podría ser cada vez más difícil identificar el trato diferencial basado en motivos protegidos en el contexto de los sistemas algorítmicos, la noción de discriminación indirecta podría convertirse en un «refugio» conceptual para cubrir los errores de los sistemas que ocasionen discriminación. Sin embargo, esta tendencia podría aumentar la inseguridad jurídica en la medida en que es más fácil justificar este tipo de discriminación. También la investigación de ADAMS-PRASSL, J. et al, «Directly Discriminatory Algorithms», *Modern Law Review*, 86(1), 2023, pp. 144–175, desafía este análisis mayoritariamente sostenido de que

conveniente que esto plantea es que es más fácil justificar la discriminación indirecta que la directa, puesto que ésta tiene unas causas concretas y tasadas de justificación²⁸⁸.

La discriminación algorítmica *directa* podría darse en dos casos: por un lado, cuando al diseñar el algoritmo se introducen datos o características protegidas de las personas a las que se asocia un peso o valor negativo, de forma que ese criterio conlleva automáticamente un resultado perjudicial para todos los miembros del grupo protegido (por ejemplo, cuando ser mujer excluya de forma automática a la hora de conceder un crédito), o contribuye a reducir la puntuación de la persona cuando dicho valor negativo se pondera con otras variables (por ejemplo, cuando ser mujer hace que disminuyan las posibilidades de obtener un préstamo, pero se pondera con otras variables para determinar el resultado final); por otro lado, cuando la pertenencia de una persona al grupo desaventajado se infiera de otros datos a los que se atribuya un valor negativo²⁸⁹. En este último caso, como explica SORIANO ARNANZ, para que se trate de discriminación directa (y no indirecta) los datos de los que se infiere la pertenencia a un grupo protegido no deben ser relevantes para predecir el aspecto de la realidad social que el algoritmo trata de medir. Es decir, las decisiones no han de tomarse en base a criterios aparentemente neutros u objetivos, sino en base a datos que, a pesar de no tener valor predictivo en la decisión que se adopta, ayudan a inferir la pertenencia de la persona al grupo protegido²⁹⁰. Así ocurriría, por ejemplo, si al predecir la solvencia de una

la discriminación algorítmica se limita a formas indirectas, argumentando que si bien la discriminación indirecta tiene un papel importante en las decisiones algorítmicas, es un error reducirlo exclusivamente a este ámbito puesto que ciertas formas de sesgo algorítmico en algoritmos frecuentemente utilizados pueden constituir discriminación directa. Pone numerosos ejemplos de sistemas que se basan directamente en características protegidas para asignar plazas en centros de estudio, generar puntuaciones crediticias o detectar fraudes en las prestaciones. Vid., pp. 157 y ss.

²⁸⁸ Por lo que se refiere a la justificación de la discriminación, si bien en el marco del CEDH la discriminación directa está sujeta a la causa de justificación general denominada justificación objetiva, en la legislación de la UE, las causas de justificación frente a la misma son limitadas. Sin embargo, por lo que se refiere a la discriminación indirecta, todas sus formas están sujetas a la causa de justificación objetiva, con independencia de que la reclamación se realice al amparo del CEDH o de la legislación de la UE.

²⁸⁹ Así lo explica SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación...», *cit.*, p. 8-9/31, clasificando la discriminación directa en «discriminación directa en sentido estricto» y discriminación «directa por inferencia» (introducida de forma consciente o inconsciente por el programador o derivada del propio aprendizaje automático del algoritmo). También en SORIANO ARNANZ, A., «La aplicación del marco jurídico europeo...», *cit.*, pp. 74-76.

²⁹⁰ SORIANO ARNANZ, A., «Discriminación algorítmica: garantías y protección jurídica», en *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, coord. por M. Bauzá Reilly; dir. por L. Cotino Hueso, 2022, p. 150, explica que el algoritmo puede inferir que una persona pertenece a un grupo protegido porque un dato es indicativo de esa condición, aunque no determinante.

mujer para la posible concesión de un préstamo, no se tomara como criterio de valoración su nómina o ingresos anuales, sino su pertenencia a una determinada asociación de mujeres o sus aptitudes para practicar un deporte en el que mayoritariamente intervienen mujeres. También en esta modalidad de discriminación directa (por inferencia), el valor negativo asociado a la categoría de datos inferida (ej. presidenta de la asociación de mujeres contra la violencia de género = rebaja automática de la puntuación del solicitante) podrá determinar automáticamente el resultado perjudicial o contribuir a él sin ser necesariamente determinante en su obtención. En cambio, si a pesar de que el sistema atribuya un valor negativo a la pertenencia de la persona al grupo protegido, la combinación con otros factores o categorías de datos lleva a que el resultado final acabe siendo favorable al interesado, no podría entenderse que existe discriminación, a pesar de que en la toma de decisión sí que existiría un trato discriminatorio²⁹¹.

Un caso de *discriminación directa por inferencia* fue el ocurrido con el algoritmo que Amazon desarrolló para el cribado y selección de currículums para puestos técnicos en la empresa. Dicho algoritmo se entrenó con los datos de las decisiones de contratación y currículums enviados a la empresa durante los diez años anteriores. A partir de dichos datos, el algoritmo debía extraer las características relevantes que debían buscarse en los currículums de los futuros solicitantes. El problema fue que la mayoría de currículums presentados en ese periodo de tiempo eran de hombres (algunos de los cuales lógicamente acabaron siendo contratados), con lo que el algoritmo dedujo que era preferible contratar hombres que mujeres y comenzó a discriminar a personas en cuyos currículums aparecían palabras de las que se podía inferir su género femenino, a pesar de que esos datos eran irrelevantes para predecir la idoneidad de los candidatos para el puesto de trabajo (ej. pertenencia a asociaciones de mujeres, haberse graduado en una universidad de mujeres, adjetivos mayoritariamente utilizados por mujeres, etc.). Aunque la variable «sexo» no aparecía en el procesamiento, el algoritmo aprendió que era una característica importante a buscar en los nuevos currículums y acabó puntuando peor a las mujeres porque aprendió a buscar e inferir la categoría sospechosa de otros datos. A nuestro modo de ver, esto es lo realmente importante del asunto. Las

Por ejemplo, haber ido a un colegio en el que la mayoría del alumnado pertenece a una minoría étnica implica una mayor probabilidad de formar parte de ese grupo, pero no es una variable totalmente predictiva de la pertenencia al grupo si el alumnado no se compone exclusivamente de miembros de dicha etnia. Otra posibilidad sería la inferencia de la pertenencia al grupo especialmente protegido a través de datos que sean inseparables a dicha pertenencia. En el ejemplo anterior, si resultara que en ese colegio únicamente acceden miembros de un determinado grupo étnico o racial, ese dato (la pertenencia a dicho colegio) sería inseparable de la pertenencia al grupo.

²⁹¹ SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación...», *cit.*, p. 9/31.

inferencias a las que llegó el algoritmo en este caso fueron fruto de su propio autoaprendizaje. Es decir, nadie había ordenado al algoritmo que discriminara por razón de sexo, sino que fue el propio sistema el que de forma autónoma llegó a tales conclusiones tras su puesta en funcionamiento²⁹².

Sin perjuicio de que puedan darse casos de discriminación directa en sentido estricto a través de sistemas de IA, compartimos la postura de quienes afirman que estos casos no serán frecuentes en la medida en que es poco probable (aunque posible) que en modelos supervisados los programadores introduzcan directamente motivos protegidos como variables a las que asociar un peso o puntuación negativa.

Son varias las razones que pueden esgrimirse para sostener lo afirmado: en primer lugar, que en las sociedades occidentales existe a todos los niveles una consolidada prohibición de discriminación de la que suele ser consciente, o habría de serlo, el ciudadano medio. Además, tal proscripción no solo viene recogida en la normativa antidiscriminatoria sino también en la normativa de protección de datos personales que ha de respetarse por quienes trabajan con ellos. Por tanto, la diligencia del buen profesional exigiría, a nuestro modo de ver, un adecuado conocimiento de la imposibilidad legal de tratar ciertas categorías de datos personales especialmente protegidos en virtud del art. 9 del RGPD, que prácticamente vienen a coincidir con las denominadas categorías sospechosas en la normativa antidiscriminatoria²⁹³. Por ello, en la medida en que los programadores de estos sistemas conozcan éstas y otras obligaciones legales dirigidas a garantizar el derecho a la igualdad de trato y a la no discriminación, será menos probable que el origen de la misma se ubique en la fase de diseño del algoritmo²⁹⁴; en segundo lugar, que el carácter especialmente

²⁹² A estos casos parece hacer referencia ESPÍN ALBA, I., *op.cit.*, p. 144, cuando habla de «sesgo algorítmico propiamente dicho», es decir, casos en que los algoritmos de IA crean sus propios sesgos a causa de la elección de características, arquitectura del modelo y parámetros, por las decisiones del propio modelo. Sobre el caso de Amazon y su calificación como discriminación algorítmica directa, vid., ADAMS-PRASSL, J. et al, *op.cit.*

²⁹³ Salvo en determinadas circunstancias mencionadas en el art. 9.2, «[q]uedan prohibidos el tratamiento de datos personales que revelen el origen étnico o racial, las opiniones políticas, las convicciones religiosas o filosóficas, o la afiliación sindical, y el tratamiento de datos genéticos, datos biométricos dirigidos a identificar de manera unívoca a una persona física, datos relativos a la salud o datos relativos a la vida sexual o la orientación sexual de una persona física».

²⁹⁴ No obstante, ha de advertirse que, en ocasiones, eliminar las características protegidas del conjunto de datos disponibles podría ser problemático ya que podría conducir a ocultar la discriminación por completo o a reducir la precisión predictiva del modelo. Y ello por varias razones: aunque se elimine una categoría protegida (como el género), otras variables correlacionadas con ella (como el puesto de trabajo ocupado, el nivel de ingresos o la ubicación) podrían actuar como *proxies*, dificultando la separación del impacto específico de la categoría protegida en los resultados del modelo; además, si se eliminan variables estrechamente correlacionadas con la variable objetivo, el modelo

reprochable que se otorga a la discriminación y que lleva a su prohibición, hace que la misma no suela manifestarse de forma abierta y claramente identificable, sino que lo más normal es que sea una actuación más presunta que patente y por ello difícil de demostrar²⁹⁵; por último, que discriminar de forma directa a través de estos sistemas carecería de sentido si se tiene en cuenta que los mismos ofrecen muchas posibilidades para ocultar la discriminación.

Expertos en la materia mantienen que el propio funcionamiento de los algoritmos, el hecho de que dependan de datos sociales que en muchos casos están sesgados *de facto*, así como la granularidad de los datos, unido a las dificultades para probar la discriminación directa si no se tiene acceso pleno al sistema²⁹⁶, podría conllevar que muchos más casos de discriminación se acabaran calificando como discriminación indirecta (en lugar de directa) con su consiguiente mayor margen de justificación²⁹⁷.

En suma, existe cierto consenso en que la discriminación indirecta encaja mejor con los modelos algorítmicos. Por tanto, en la mayoría de casos habrá

puede perder información relevante para hacer predicciones precisas. Sobre esta cuestión, HACKER, P., «Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law», 55 *Common Market Law Review*, 2018, pp. 19-20. Considera este autor, que reducir la precisión predictiva de un modelo podría ser aceptable si se logra minimizar la discriminación, pero que esta decisión dependerá del contexto y la proporcionalidad del uso del modelo. Es decir, las decisiones algorítmicas que afectan a grupos protegidos solo pueden justificarse si cumplen con pruebas de proporcionalidad, lo que incluye la evaluación de alternativas menos discriminatorias. Además, en su opinión, la valoración de los costes y beneficios que conllevaría reducir la precisión del modelo ha de hacerse caso por caso. Por su parte, NAVAS NAVARRO, S., «La perspectiva de género en la inteligencia artificial», *Diario La Ley*, n° 48, Sección Ciberderecho, 8 de marzo de 2021, p. 12/22, apunta que por ejemplo en el ámbito de la salud es imprescindible no obviar el sexo o el género por ser una variable fundamental para llegar a soluciones acertadas y precisas. Para llegar a resultados justos, no considera que sea adecuada la solución propuesta de eliminar del sistema aquellos datos sensibles como puede ser el sexo o el género. Y ello porque, además del hecho de que desagregar los datos con base en estos indicadores es un mandato legal, en no pocos ámbitos, como por ejemplo en medicina, la presencia de estos datos es esencial. De hecho, el uso explícito de estos indicadores conduce a soluciones más justas en la medida en que pueden apreciarse las diferencias de manera que se puedan proponer soluciones adecuadas y precisas para cada individuo. Por tanto, lo que propone esta autora para mitigar la existencia del sesgo de sexo y género es proporcionar al sistema una instrucción explícita que incluya reglas acerca del comportamiento más apropiado.

²⁹⁵ TOLOSA, P./DIBO, C., *op.cit.*, p. 159.

²⁹⁶ Da amplia cuenta de este hándicap, SORIANO ARNANZ, A., «La aplicación del marco jurídico europeo...», *cit.*, p. 74. Según señala la autora, la falta de acceso a un sistema algorítmico puede derivar en que no se pueda demostrar un caso de discriminación directa. E incluso aunque se tenga acceso, si el sistema resulta tan incomprensible que no se pueden identificar con claridad las variables introducidas, será complicado detectar si la categoría sospechosa se tiene en cuenta de manera explícita o no. Por ello, será muy probable que muchos casos de discriminación algorítmica sean tratados como discriminación indirecta.

²⁹⁷ GERARDS, J./XENIDIS, R., *op.cit.*, p. 72.

que centrarse en el impacto, efectos o consecuencias del algoritmo. La *disposición, criterio o práctica aparentemente neutros* que impactan negativamente en un grupo protegido puede ser *alguna variable* introducida en el sistema que, por el valor que éste le otorga, genera un resultado discriminatorio para el grupo desaventajado²⁹⁸.

Un caso en que se apreció discriminación algorítmica indirecta fue el del algoritmo «Frank» empleado por la empresa Deliveroo. Tal como se explicó en su momento, el Tribunal Ordinario Bolonia en su sentencia de 31 de diciembre de 2020, consideró que el algoritmo de clasificación utilizado por la plataforma digital para la asignación y distribución de pedidos a sus repartidores era discriminatorio. Los criterios aparentemente neutros que tenía en cuenta la empresa para llevar a cabo dicho ranking eran la *fiabilidad* y *disponibilidad* del repartidor, e implicaban que cuantas más horas estuviera disponible el trabajador y menos cancelaciones de franjas horarias previamente reservadas tuviera, mejor posicionamiento le asignaba el sistema, lo que a su vez implicaba que la próxima vez que accediera a la plataforma tuviera preferencia para elegir el horario que más le conviniera.

El órgano judicial mantuvo que, a pesar de que la finalidad perseguida por la empresa era legítima en tanto que pretendía fomentar el trabajo en horas de alta demanda y evitar cancelaciones de franjas horarias previamente reservadas, el algoritmo *discriminaba indirectamente* a ciertos trabajadores en las condiciones de acceso al trabajo porque no tenía en cuenta, para llevar a cabo el ranking reputacional, los motivos por los que el trabajador había tenido que cancelar las sesiones de trabajo previamente reservadas, penalizando de la misma forma al que cancelaba su reserva por motivos justificados (ej. enfermedad, cuidado de menores o ejercicio del derecho a la huelga, entre otros), y al que lo hacía por motivos no justificados, lo que no se ajustaba al Derecho laboral. Por tanto, al tratar igual a personas que no se encontraban en análoga situación (por cuanto que unas causas de absentismo laboral están amparadas por la legislación y otras no), se las estaba discriminando injustamente²⁹⁹. A mayor abundamiento, la sentencia declara que las instrucciones que la empresa proporcionó al algoritmo ignoraban estos motivos de forma *consciente*.

Ahora bien, a diferencia de lo ocurrido en el caso Deliveroo donde esos criterios aparentemente neutros eran públicos (ya que aparecían en su página web como condiciones de los contratos de los repartidores), en muchas ocasiones, la opacidad algorítmica complica –o incluso impide– determinar qué variable o variables o qué inferencias pueden haber ocasionado el resultado

²⁹⁸ SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación...», *cit.*, p. 11/31.

²⁹⁹ Sobre este caso nos remitimos a lo ya comentado en el apartado 1.2.1.3 y a la doctrina allí citada.

discriminatorio (es decir, impide saber la causa de la discriminación y si ésta es directa o indirecta). En estos casos, la solución pasaría por considerar al propio *algoritmo en sí mismo* como disposición, práctica o criterio aparentemente neutro³⁰⁰, siempre y cuando se consiguiera demostrar su impacto discriminatorio, para lo cual serían especialmente importantes las pruebas estadísticas³⁰¹.

Dada la gran cantidad de datos procesados por los sistemas de IA y su capacidad para identificar patrones o correlaciones entre ellos, también podría ocurrir que el algoritmo clasificara erróneamente a una persona, bien por su vinculación con otra, bien por equivocación, dando lugar a *discriminaciones por asociación o error*, respectivamente. Por ejemplo, si clasifica a un usuario dentro de una categoría protegida sin que en él concurra dicha característica, sino por su proximidad o relación con otra persona o grupo que ostenta el rasgo protegido, podría darse una discriminación por asociación. A dicha asociación podría llegar, por ejemplo, mediante la recopilación de ciertos datos de la persona como el historial de búsquedas en Internet, los *clicks* a los que da o los «me gusta» que pone en redes sociales³⁰². Dichos datos pueden indicar que el usuario está interesado en saber si determinados establecimientos son accesibles para personas con discapacidad motora y en base a esto el algoritmo podría perfilarlo como «persona con discapacidad», cuando en realidad no lo es, pero comparte su vida con alguien que sí lo es. Consecuentemente, podría dejar de mostrarle ciertas ofertas de trabajo o incluso excluirla de ciertos procesos de selección de personal en función de esa clasificación o perfilado incorrecto hecho por el algoritmo.

La doctrina considera que no cabe identificar *asociación* con *correlación de datos de proxy*, ya que técnicamente el *proxy* (o variable de aproximación) en base al cual el algoritmo toma su decisión, no constituye en sí mismo una causa de discriminación prohibida, sino que es un elemento que detecta características por proximidad. Es decir, tomar decisiones en base, por ejemplo, al acceso de una persona a determinadas webs o foros digitales no resulta *a priori*

³⁰⁰ SORIANO ARNANZ, A., «La aplicación del marco jurídico europeo...», *cit.*, p. 76.

³⁰¹ Conviene recordar la importancia de la evidencia estadística como indicio de discriminación indirecta para aplicar la inversión de la carga de la prueba (STJUE 28.2.2013 (caso c-427/11, asunto *Kenny et al.*). Recoge el *Manual de legislación europea contra la discriminación*, *cit.*, p. 63, que al analizar las pruebas estadísticas del efecto desproporcionadamente perjudicial para el «grupo protegido» respecto a otras personas en situación similar, el TJUE y el TEDH buscan muestras de que una proporción especialmente grande de los perjudicados pertenezca a dicho grupo, pero sin fijar un porcentaje concreto y analizando cada situación en particular. La relevancia de este tipo de prueba estadística usada en los casos de discriminación por sistemas de IA, tanto en Europa como Norteamérica la destaca entre los autores MUÑOZ GUTIÉRREZ, C., *op.cit.*, p. 289.

³⁰² ORWART p. 34, pone como ejemplo el sistema empleado por Facebook para excluir del acceso a publicidad sobre venta y alquiler de inmuebles y otros productos, a personas relacionadas con determinados grupos raciales, personas con discapacidad y otros grupos protegidos.

una práctica discriminatoria al no ser ésta una causa de discriminación prohibida. Tampoco sería arbitraria atendiendo a la correlación identificada por el algoritmo. Quizá por ello hay quien afirma que no es posible mantener un concepto «digital» de discriminación por asociación a pesar de la conexión que existe entre ambos conceptos y del elevado potencial discriminatorio que tienen estos sistemas por su precisión a la hora de encontrar patrones entre los datos³⁰³.

Ahora bien, a nuestro modo de ver, eso no significa que si las correlaciones a las que llega el algoritmo a partir de esos *proxies* hacen que el sistema (o sus resultados) acabe impactando negativamente en una persona o grupo protegido, no puedan dar lugar a discriminaciones ilícitas. Para GINÉS I FÀBREGAS «el hecho que la decisión se haya basado en una variable *proxy* o en una inferencia realizada por el algoritmo es irrelevante jurídicamente, por cuanto aquello relevante es la existencia de un efecto discriminatorio. Así, ya sea porque la variable *proxy* es el reflejo de una variable de discriminación prohibida o porque, a pesar de ser una variable aparentemente neutra e inofensiva, genera una desventaja particular a un grupo protegido, resulta una práctica discriminatoria jurídicamente prohibida»³⁰⁴.

Junto con la discriminación por asociación, es posible también que el algoritmo discrimine por error, esto es, porque considera que la persona ostenta ciertas características que en realidad no tiene. Se diferencia del caso anterior, en que aquí no hay otro individuo o grupo especialmente vinculado con la persona objeto de discriminación que sí ostenta la característica protegida, sino que se trata simplemente de que, por una apreciación incorrecta de sus características o su comportamiento, el algoritmo adscribe a la persona a un determinado perfil o le adjudica cierto rasgo que no le corresponde. Es decir, en atención a las características o aptitudes que presentan personas similares,

³⁰³ En este sentido, RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», *cit.*, p. 423, quien no tiene claro cómo interpretarán los tribunales estos casos de sesgos o discriminación por inferencias o correlaciones algorítmicas. Entre otros muchos autores, ORWART, C., *op.cit.*, trata ampliamente el tema de los *proxies* y los problemas de discriminación en base a ellos con muchos ejemplos de casos reales. Y en la misma línea advierte MUÑOZ GUTIÉRREZ, C., *op.cit.*, p. 281, del peligro que conllevan los *proxies* ya que, bien de forma accidental o bien de forma intencional, pueden corresponderse con marcadores de características protegidas. Es decir, determinados *clicks* pueden marcar una orientación sexual erróneamente, así como un código postal puede marcar una nacionalidad, condición social o etnia equivocada. Por tanto, además de ser imprecisos, el uso de *proxies* puede conllevar arbitrariedades y discriminación.

³⁰⁴ *Op.cit.*, pp. 313-314. No obstante, para HUERGO LORA el hecho de que el algoritmo tome decisiones basadas en *proxies* sí es jurídicamente relevante, ya que eso podría suponer que el algoritmo estuviera utilizando un hecho o circunstancia distinta a la prevista en la norma para aplicar una consecuencia jurídica. Es decir, estaría sustituyendo el supuesto de hecho previsto en la norma por otro distinto correlacionado con él.

el algoritmo hace una estimación (equivocada) sobre una persona concreta, de forma que existe discriminación cuando la persona que no forma parte del grupo protegido ha sido discriminada a consecuencia de haber sido erróneamente clasificada dentro de él³⁰⁵. Este tipo de discriminación también cobra una particular relevancia en el ámbito de los algoritmos porque en muchos casos la elaboración de perfiles dará como resultado la atribución de determinados rasgos o características a personas o grupos basándose, por ejemplo, en sus datos de comportamiento en redes sociales y tales inferencias pueden ser incorrectas (errores que pueden deberse, por ejemplo, a la insuficiencia o falta de datos, falta de granularidad, etc.).

El TJUE no se ha pronunciado de forma consistente acerca de si la discriminación por error puede recaer bajo el concepto de la discriminación directa. No obstante, teniendo en cuenta que la discriminación por error se origina porque el algoritmo *asocia* erróneamente a una persona con un grupo protegido, un sector doctrinal ha defendido la posibilidad de aplicar la doctrina sentada por el TJUE en el *caso Coleman* (donde se apreció discriminación por asociación y se consideró como una discriminación directa), también a los casos de discriminación algorítmica por error³⁰⁶. La Comisión Europea también defiende que la discriminación por error pueda englobarse bajo el paraguas de la discriminación directa teniendo en cuenta que tanto la Directiva sobre igualdad racial, como la Directiva sobre igualdad de trato en el empleo y la ocupación, prohíben una situación en la que una persona sea *directamente* discriminada sobre la base de una percepción o suposición errónea de características protegidas. Si la discriminación directa no englobara tales situaciones, podría ser difícil calificar aquellos casos en que las identidades atribuidas por el algoritmo conllevaran un trato diferencial y desventajoso para personas o grupos³⁰⁷.

Por último, se ha demostrado que en la toma de decisiones algorítmicas son particularmente generalizadas las *formas interseccionales de discriminación* y que el uso de los sistemas de IA aumentará considerablemente el riesgo de que se den discriminaciones múltiples e interseccionales³⁰⁸. Esto es así porque

³⁰⁵ No obstante, debe tenerse en cuenta que estos sistemas y los perfilados que realizan cuentan generalmente con un porcentaje de error; es decir, personas erróneamente clasificadas en un perfil (falsos positivos) y personas erróneamente excluidas del mismo (falsos negativos).

³⁰⁶ En tal sentido se posiciona GINÈS i FABRELLAS, A., *op.cit.*, p. 315. Pero en cualquier caso, como indica la autora, «la existencia de discriminación por error requeriría acceder, por un lado, a información sobre el perfil elaborado de la persona, para confirmar la errónea clasificación de la persona, como, por otro lado, a la información estadística sobre el impacto discriminatorio del algoritmo».

³⁰⁷ Así lo consideran los autores del estudio *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law*, ya citado, pp. 68-69.

³⁰⁸ GERARDS, J./XENIDIS, R., *op.cit.*, pp. 65-66 y p. 142.

la cantidad y granularidad de los datos que manejan estos sistemas de automatización de decisiones hace que sea inusual que el algoritmo discrimine exclusivamente o de manera decisiva en base a un único criterio protegido, ya que generalmente basará su resultado en una multitud de factores y variables diferentes estadísticamente correlacionadas. En consecuencia, en la mayoría de ocasiones el resultado de un algoritmo o la decisión adoptada en atención a él no se basarán *únicamente* en el sexo, la etnia o la discapacidad de una persona, sino en una combinación de esas características y comportamiento que es exclusivo de una persona o grupo, no siendo ni siquiera posible identificar cuál de estas características tuvo más peso para la adopción de la decisión³⁰⁹.

Para que los daños ocasionados por este tipo de discriminación puedan ser reparados de forma efectiva es preciso un mayor reconocimiento legal de dicha modalidad, no sólo a nivel nacional sino también europeo, tarea que en nuestro territorio ha sido cumplida en fechas recientes a través de la promulgación de la Ley 15/2022, como se comentó anteriormente.

Sin perjuicio de lo anterior, debe advertirse que ciertas características intrínsecas a estos sistemas decisionales –como su opacidad y complejidad técnica–, implican problemas para clasificar los casos de discriminación algorítmica dentro de una u otra tipología al ser muy difícil identificar cuál es su causa u origen. Es decir, será difícil (cuando no imposible) saber si se basa o no en motivos protegidos a efectos de poder considerarla como directa, especialmente por la necesidad de aportar una referencia comparativa. Por tanto, si la falta de transparencia sobre el funcionamiento del algoritmo impide recabar pruebas acerca de cómo ha tratado o habría tratado el algoritmo a un grupo que no comparte la característica protegida en juego (el grupo de comparación), entonces sería imposible acreditar la discriminación³¹⁰. Por todo ello se corre el riesgo de que cualquier caso en que el algoritmo arroje resultados discriminatorios se considere como *discriminación indirecta*, lo que aumentaría la inseguridad jurídica para posibles víctimas, desarrolladores y usuarios de sistemas algorítmicos por cuanto que dicha forma o modalidad discriminatoria admite un mayor margen de justificación que la directa³¹¹.

En consecuencia, y a modo de conclusión, se puede afirmar que el uso de sistemas de IA en la toma de decisiones ha reconfigurado hasta cierto punto la división tradicional entre la discriminación directa e indirecta.

³⁰⁹ Un ejemplo de discriminación interseccional podría ser el de los sistemas de reconocimiento facial basados en imágenes en el que las mujeres de piel oscura experimentaban un trato peor por más de un motivo (en este caso, ser mujer y además de raza negra). Nos referiremos a estos casos, *infra*.

³¹⁰ Sobre los problemas de prueba hablaremos más adelante.

³¹¹ GERARDS, J./XENIDIS, R., *op.cit.*, pp. 75 y ss.

3. CAUSAS POR LAS QUE SE PUEDE PRODUCIR LA DISCRIMINACIÓN ALGORÍTMICA

3.1. CONSIDERACIONES PREVIAS

Para hacer frente al problema de la discriminación algorítmica se torna imprescindible examinar los motivos o causas por los que un sistema inteligente puede llegar a resultados discriminatorios, ya que sólo detectando sus causas se podrán evitar, o al menos minimizar, sus consecuencias.

Tal como advertía la Comisión en el *Libro Blanco sobre IA*, los riesgos que puede conllevar el uso de esta tecnología para los derechos fundamentales pueden deberse tanto a defectos en el *diseño* de los sistemas –especialmente en lo que se refiere a la supervisión humana–, como al uso de *datos* que puedan estar sesgados al no haber sido previamente corregidos, lo que podría ocurrir, por ejemplo, cuando se entrena un sistema utilizando única o principalmente datos relativos a hombres y ello se traduce en peores resultados para las mujeres³¹².

También el Parlamento Europeo, en la Propuesta de Reglamento que acompañaba su *Resolución sobre aspectos éticos de 20 de octubre de 2020*, establecía que el sesgo podía originarse tanto en las *decisiones* basadas en un sistema automatizado o adoptadas por tal sistema, como en los conjuntos de *datos* en los que se basan dichas decisiones o con los que se entrena al sistema³¹³.

Y con algo más de detalle afirmaba el Relator especial de la ONU sobre la promoción y protección del derecho a la libertad de opinión y de expresión, en su *Informe sobre las Tecnologías de inteligencia artificial e implicaciones para el entorno de la información*, que existen diversos factores que incorporan sesgo en los sistemas de IA y aumentan su potencial discriminatorio, entre los cuales menciona la forma en que se *diseñan*, los *conjuntos de datos con que se entrenan*, los *sesgos sociales y culturales que los creadores* de aplicaciones pueden incorporar en esos conjuntos de datos, los propios *modelos de IA* y la *forma en que los productos del modelo de IA se ejecutan* en la práctica³¹⁴.

No cabe duda, por tanto, de que son múltiples los factores que pueden motivar el sesgo en los sistemas de aprendizaje automático. Aunque los expertos mantienen que en la mayor parte de los casos se debe a cuestiones

³¹² *Libro Blanco sobre IA*, cit., p. 13.

³¹³ Considerando 22.

³¹⁴ A/73/348 p.15/ 24. Recuperado de: <https://www.ohchr.org/es/special-procedures/sr-freedom-of-opinion-and-expression/annual-thematic-reports>.

relacionadas con los datos utilizados para entrenar el sistema³¹⁵, la discriminación algorítmica no es sólo problema de los datos, sino que incluso cuando estos son adecuados, el algoritmo puede llevar a resultados injustos ya que van aprendiendo de su propia experiencia y decidiendo en base a ella, sin que ni siquiera en algunos casos el programador pueda explicar el porqué de tal decisión³¹⁶. Se mantiene, por ello, que los sesgos pueden provenir también de la llamada *caja negra*, es decir, de la definición del algoritmo, de la forma en que éste trabaja y de su facultad de predicción³¹⁷.

También es importante advertir que el origen de la discriminación puede ubicarse en las diferentes etapas del proceso de toma de decisión algorítmica, esto es, tanto en la *fase de planificación*, donde se decide si se va a utilizar un algoritmo para la toma de decisiones, qué tipo de algoritmo se utilizará –por ejemplo, basado en reglas o aprendizaje automatizado–, cómo se utilizará –si para adoptar de forma definitiva decisiones o como ayuda para la toma de decisiones humanas–, quién va a desarrollar el algoritmo, qué conjuntos de datos se utilizarán para entrenar el algoritmo, etc., como en la *fase de desarrollo*, que se refiere principalmente a la creación y entrenamiento del algoritmo, como en la propia *fase de aplicación y uso*, esto es, a la implementación del algoritmo y

³¹⁵ Así lo afirman la gran mayoría de expertos que tratan el tema tanto desde el punto de vista técnico como jurídico. Por ejemplo, BAROCAS, S./SELBST, A. D., *op.cit.*, o HACKER, P., *op.cit.*, pp. 5-7, para quien el sesgo en los datos de entrenamiento es un problema crítico y gran parte de la discriminación algorítmica se deriva de datos de entrenamiento sesgados que reflejan prejuicios históricos o errores de muestreo, lo que perpetúa patrones de discriminación en los resultados algorítmicos. Para SOLAR CAYÓN, J.I., *op.cit.*, p. 157, aunque se hable generalmente de «sesgo algorítmico», la parcialidad de los resultados de los sistemas de aprendizaje automático se debe en la mayoría de casos a los datos de entrenamiento del sistema. Puesto que el algoritmo aprende de estos, los errores y prejuicios en ellos contenidos, contaminarán ineludiblemente sus resultados, pudiendo generar discriminación. ESPÍN ALBA, I., *op.cit.*, p.132, apunta que «la esencia técnica del problema está en gran medida –si bien no exclusivamente– en la selección, calidad y tratamiento de los datos de entrada (*input*), así como en la interpretación de los resultados (*output*). Por lo tanto, la primera respuesta son los parámetros de transparencia, rendición de cuentas y responsabilidad ensayados en el régimen europeo de protección de datos». RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», en ÁLVAREZ LATA, N. (coord.), *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial (APPDC)*, Madrid, 2024, p. 664, confirma que es muy numerosa la literatura que mantiene el origen de los sesgos en la calidad de los datos utilizados inicialmente para entrenar al algoritmo. También organismos como la AEPD considera que, en el caso de tratamientos de datos que incorporan soluciones IA, la existencia de sesgos en los modelos de inferencia está íntimamente ligada con la exactitud o calidad del dato. Vid., *Informe sobre la Adecuación del RGPD a la IA*, cit., p. 35.

³¹⁶ Esto puede ocurrir con los algoritmos de aprendizaje automatizado, sobre todo cuando se trata de aprendizaje no supervisado y aprendizaje profundo. Por ello, en estos casos, habrá que ir monitorizando el algoritmo para comprobar si sigue generando resultados fiables y aceptables. Así se explica en GERARDS, J./ XENIDIS, R., *op.cit.*, p. 40.

³¹⁷ SÁEZ LARA, C., *op.cit.*, p. 46.

al control del sistema tras su puesta en práctica para asegurar que, aun a pesar de sus adaptaciones, sigue generando decisiones fiables y aceptables³¹⁸.

3.2. FACTORES QUE PUEDEN MOTIVAR LA DISCRIMINACIÓN DE LOS SISTEMAS DE IA

Entre las causas o factores que pueden motivar el sesgo en los sistemas inteligentes para la toma de decisiones se pueden mencionar, sin ánimo de exhaustividad, los siguientes³¹⁹:

³¹⁸ Sobre cada una de estas fases puede verse GERARDS, J./ XENIDIS, R., «Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law» (Special report European network of legal experts in gender equality and non-discrimination), European Commission, 2020., p. 37. También explica el ciclo de vida de los sistemas de IA y sus fases, PALMA ORTIGOSA, A., «El ciclo de vida de los sistemas de inteligencia artificial. Aproximación técnica de las fases presentes durante el diseño y despliegue de los sistemas algorítmicos», en *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, M. Bauzá Reilly (coord.); L. Cotino Hueso (dir.), Navarra, 2022, pp. 29-51. En la *Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno* del Consejo de Europa se advierte que «Se debe tener especial cuidado en las fases de desarrollo e implementación, especialmente cuando el procesamiento es directo o indirectamente basado en datos “sensibles”. Esto podría incluir supuesto origen racial o étnico, antecedentes socioeconómicos, opiniones políticas, creencias religiosas o filosóficas, afiliación sindical, datos genéticos, datos biométricos, datos relacionados con la salud o datos relacionados con la vida sexual u orientación sexual».

³¹⁹ Mientras hay autores que reducen a dos estas causas (a saber, diseño del algoritmo y *data set* de entrenamiento) –así, SOLAR CAYÓN, J.I., *op.cit.*, p. 157 o CASTILLO OCARANZA, C., *op.cit.*, p. 76–, otros hacen una clasificación un poco más amplia: vid., por ejemplo, EGUÍLUZ CASTAÑEIRA, J.U., *op.cit.*, p. 345 y ss; BAROCAS, S./SELBST, A. D., *op.cit.*, pp. 677 y ss, identifican como posibles causas la definición de las variables objetivo y las etiquetas de clase, puesto que las elecciones subjetivas en esta etapa pueden introducir sesgos; la selección de características, puesto que la omisión de factores relevantes puede llevar a clasificaciones inexactas, especialmente para grupos protegidos; y los datos de entrenamiento sesgados o no representativos, que pueden reflejar discriminación previa y perpetuar prejuicios. TOLOSA, P./DIBO, C., *op.cit.*, p. 179; SORIANO ARNANZ, A./SIMÓ SOLER, E., *op.cit.*, pp. 194 y ss; GONZÁLEZ VALVERDE, A., «Responsabilidad por el potencial de sesgo discriminatorio de los algoritmos de los productos de Inteligencia Artificial. A propósito de la Algorithmic Accountability Act of 2019 (s.1108)», en Ataz López y Cobacho Gómez, *Cuestiones clásicas y actuales del Derecho de daños: Estudios en homenaje al profesor Dr. Roca Guillamón*, Vol. 2, 2021, p. 1232, apunta tres problemas habituales en las herramientas de aprendizaje automático: 1) Distribuir los datos en conjuntos de prueba y de entrenamiento de forma inapropiada, esto es, al azar, al crear modelos ignorando que en la vida real los datos rara vez son aleatorios; 2) ignorar la existencia de variables ocultas e incontroladas; 3) y equivocar el objetivo, al utilizar un número muy reducido de «funciones de pérdida»; GINÈS i FABRELLAS, A., *op.cit.*, p.302, menciona cuatro factores o una combinación de ellos: (1) discriminación por sesgos en las variables utilizadas por el algoritmo para tomar decisiones, (2) discriminación por sesgos en los datos sobre los que se ha entrenado el algoritmo, (3) discriminación por correlaciones realizadas por el algoritmo o (4) discriminación por el efecto que generan las decisiones adoptadas mediante algoritmos.

3.2.1. El factor humano

Numerosos estudios demuestran que los algoritmos no son instrumentos infalibles de las matemáticas que producen siempre y en todo caso resultados objetivos, racionales e imparciales, sino que son tan «defectuosos» como los propios humanos que los desarrollan y los datos con los que se les entrena³²⁰. Esto es, los algoritmos, al ser diseñados y entrenados por personas pueden heredar tanto los sesgos o prejuicios de éstas, como los que están intrínsecos en los propios datos que utilizan (sesgos sociales y culturales)³²¹. Además, las personas se equivocan y los algoritmos fallan, pudiendo deberse esos errores no sólo a un mal diseño o programación que lleve al algoritmo a ejecutar mal la tarea para la que fue creado, sino también a su implementación en ámbitos en los que se carece de estándares consistentes o de una incorrecta interpretación de sus resultados³²².

La AEPD ha puesto de relieve algunos de estos fallos en la implementación de los sistemas. Sobre esto considera que «existen IA's, como por ejemplo sistemas expertos basados en reglas, en los que la concepción del propio sistema puede introducir errores que conduzcan a inferencias erróneas; o sistemas basados en ML que no son capaces de modelar el tratamiento deseado. Se pueden producir errores porque elementos externos a la IA, como por ejemplo lectores biométricos, introduzcan errores en los datos de entrada. Por otro lado, existe la posibilidad de que existan errores de programación o

³²⁰ Prácticamente todos los autores que han tratado el tema destacan esta idea. Por ejemplo, O'NEIL, C., *op.cit.*, desmonta por completo esta idea en su monografía. Y también EUBANKS, V., *op.cit.*, explica que se suele pensar que los algoritmos son neutros, pero no es así, ya que los sesgos son estructurales y sistémicos y tienen poco que ver con una decisión individual. Sobre esta idea también inciden, SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», cit., p. 93, o MUÑOZ GUTIÉRREZ, C., *op.cit.*, p. 282. RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», cit., p. 436, habla de «apariencia de objetividad» o «turbia ilusión de objetividad». También GINÉS i FABRELLAS, A., *op.cit.*, p. 321, considera importante «deconstruir el discurso de los algoritmos como modelos estadísticos puramente neutros y matemáticamente objetivos y aceptar la realidad sesgada y discriminatoria de la que se han nutrido».

³²¹ BAROCAS, S./SELBST, A. D., *op.cit.*, p. 691, mantienen que la mayoría de veces el impacto dispar o discriminatorio del algoritmo no es buscado de forma consciente, ni se debe a las creencias prejuiciosas de los que toman decisiones, sino que es involuntario y se debe a la desigualdad existente en la sociedad.

³²² El documento del NIST (*National Institute of Standards and Technology* o Instituto Nacional de Estándares y Tecnología de los Estados Unidos) titulado «Towards a Standard for Identifying and Managing Bias in Artificial Intelligence», reconoce la influencia de los sesgos humanos en el desarrollo y despliegue de sistemas de IA y destaca que los prejuicios cognitivos que las personas o grupos que diseñan y desarrollan los sistemas de IA aportan al proceso, afectan tanto a la elección de los datos que deben utilizarse, a la elección de qué modelos de IA deben desarrollarse, dónde debe implementarse el sistema de IA, o si el mismo es necesario o no.

de diseño que trasladen, de forma errónea, el modelo a su implementación práctica. En estos casos se puede decir que el que sesgo está “hardwired” por las decisiones que se toman cuando se construye el modelo de análisis (sesgos de evaluación y agregación)». Y respecto a los errores o sesgos en los conjuntos de datos de entrenamiento o validación que imposibilitan que las inferencias sean correctas, señala que los sesgos podrían ser inherentes (si ya existen en los datos que alimentan el sistema, como mala calidad de los datos, datos ausentes, o muestreo selectivo) o de representación y medida, debidos a cómo se da formato al conjunto de datos para alimentar al sistema. Por último señala que, en el caso de sistemas adaptativos, la solución IA puede ser utilizada mayoritariamente por un grupo de sujetos con características particulares que introducen nuevos sesgos de realimentación³²³.

3.2.1.1. *En el diseño del algoritmo o la selección de los datos*

Los prejuicios y estereotipos humanos pueden trasladarse al algoritmo tanto a la hora de *diseñarlo* (es decir, de escribir las instrucciones, definir la variable objetivo, introducir variables y su correspondiente peso, dividir o etiquetar los datos por categorías, etc.), como de su posterior *entrenamiento* siempre y cuando los datos que se elijan o recopilen para llevar a cabo dicha tarea no sean completos y de calidad³²⁴.

Además, esos sesgos introducidos por los programadores pueden ser tanto *conscientes* como *inconscientes*, dando lugar a discriminaciones intencionadas (*disparate treatment*) o no intencionadas (*disparate impact*). Los sistemas de IA normalmente reproducen esta última, ya que los sesgos suelen introducirse de forma involuntaria o inconsciente, lo que

a su vez conlleva que sea más difícil controlarlos y mitigarlos. En cualquier caso, en la medida en que la intencionalidad (o su falta) es irrelevante para apreciar la discriminación, podría estimarse su concurrencia sin entrar a valorar dicho extremo, es decir, con independencia de si fueron los programadores los que introdujeron los sesgos que provocaron las decisiones injustas (bien consciente o inconscientemente, bien por su propia iniciativa o siguiendo instrucciones de otra persona), o si fue el propio sistema el que las desarrolló, incluso cuando su diseño y los datos utilizados carecieran de sesgos³²⁵.

³²³ Informe sobre la *Adecuación del RGPD a la IA*, cit., p. 36/53.

³²⁴ Habría que preguntarse, como después se hará, la responsabilidad que podría derivarse para el programador o entrenador del sistema por un diseño incorrecto o una mala elección de los datos que acabe ocasionando discriminación (y daños derivados de la misma) a personas o grupos concretos.

³²⁵ GERARDS, J./XENIDIS, R., *op.cit.*, pp. 67 y ss, explican que, a tenor de la legislación de la UE, la intencionalidad del autor es irrelevante a la hora de apreciar la concurrencia de la discrimi-

Ahora bien, el hecho de que los sesgos se introduzcan dolosa o intencionadamente en el algoritmo (por ej. se diseña con el objetivo de discriminar a ciertas personas o grupos) habrá de influir en la fijación de la responsabilidad por los daños que se ocasionen.

Un caso en que el origen de la discriminación se debió a un mal diseño del sistema fue el de un algoritmo utilizado en EEUU para analizar los riesgos para la salud de millones de pacientes con el fin de asignar recursos o atención médica en función de sus necesidades. Tras una investigación se concluyó que dicho sistema discriminaba sistemáticamente a la población negra³²⁶. Pero lo llamativo del caso fue que el algoritmo ni siquiera sabía la raza de los pacientes, puesto que este dato no se le había proporcionado. ¿Cómo discriminaba si esta variable no se había incluido? Los científicos descubrieron que el fallo estaba en la «variable objetivo» de ese algoritmo, esto es, en el fenómeno que el algoritmo intentaba predecir. Era esta variable la que contenía el sesgo³²⁷.

nación directa o indirecta. Por tanto, para entender probado un caso de discriminación directa no se exige demostrar que el discriminador era consciente de la discriminación, ni demostrar que tenía intención de discriminar. Así, habría discriminación directa cuando el algoritmo se diseña de tal forma que llega a resultados discriminatorios, aun cuando esa no fuera la intención del programador, sino simplemente el resultado de sus sesgos inconscientes o de los sesgos inherentes a los datos que seleccionó para entrenar al sistema. En ese caso, bastaría probar que la categoría protegida fue tenida en cuenta por el algoritmo y en función de dicho tratamiento se derivó una desventaja para el grupo protegido, con independencia de si los desarrolladores del algoritmo, la empresa que lo utilizó con fines comerciales o la Administración que confió en él para la toma de decisiones, tenían o no intención de discriminar. Similarmente, CASTILLO OCARANZA, C., *op.cit.*, p. 73, mantiene que no es necesario que el algoritmo tenga «subjetividades» para que exista discriminación, sino sólo que arroje resultados que pongan en desventaja a un grupo protegido. Es decir, no sería necesario que el algoritmo (o más bien sus programadores) estuviera motivado por animosidad contra la gente en la que concurre cierta característica protegida. En otras palabras, la discriminación algorítmica no tiene que ver con las creencias o actitudes discriminatorias que puedan tener los desarrolladores del sistema. Y como apunta RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», *cit.*, p. 420, el hecho de que sea indiferente para determinar si existe discriminación que la misma se haya ocasionado de forma consciente y deliberada o no, permite abarcar todos los efectos derivados del sesgo algorítmico, hayan sido buscados o introducidos intencionadamente en el algoritmo, o sean espontáneos o accidentales, esto es, derivados de las inferencias de datos a las que llega el sistema.

³²⁶ El estudio puede consultarse en <https://www.science.org/doi/abs/10.1126/science.aax2342#con1>. Otro ejemplo de discriminación algorítmica que se debió a un mal diseño del sistema algorítmico fue el del algoritmo Frank, utilizado por la empresa Deliveroo, al que ya hemos hecho referencia anteriormente. En dicho caso, y a juicio del tribunal que resolvió el asunto, las instrucciones que conscientemente se le habían dado al sistema discriminaban indirectamente a ciertos trabajadores.

³²⁷ Así lo explica el principal autor del estudio, Ziad Obermeyer, de la Universidad de Berkeley (EEUU), quien afirma que la variable más importante para un algoritmo no es ninguno de los *inputs*, ni siquiera la raza. Es el *output*: la variable que el algoritmo intenta predecir. Según afirma también CASTILLA, K., *op.cit.*, p. 25, para observar si puede o no haber incidencia en la igualdad y la no discriminación, antes que en la entrada y en el procesamiento, se debe poner atención en la salida; el primer paso es determinar el objetivo del algoritmo o el resultado que se quiere obtener, ya que solo

En este caso, la predicción no eran las necesidades médicas de los pacientes sino los costes sanitarios («cuánto dinero costará un paciente»), entendiendo el programa que a mayor coste sanitario las necesidades del paciente también eran mayores. Era aquí donde entraba en juego el sesgo, al utilizar el algoritmo los costes de salud como un indicador de las necesidades médicas.

El problema fue que en EEUU los datos revelan que se gasta menos dinero en pacientes negros que tienen las mismas necesidades médicas o problemas de salud que los blancos (de media, 1800 dólares menos). Por ello, el algoritmo concluyó, erróneamente, que los negros estaban más sanos que los blancos y en consecuencia necesitaban menos recursos o atención médica (a pesar de que, según los datos, las personas de raza negra están considerablemente más enfermas que las de raza blanca)³²⁸. Según el estudio, la solución pasaba por reformular el algoritmo para que no utilizara el coste sanitario como sustituto de las necesidades de salud, eliminando así el sesgo racial a la hora de identificar quién necesitaba más ayuda médica en función de sus problemas de salud. Se concluyó que remediando esta disparidad el porcentaje de pacientes de raza negra que recibirían ayuda aumentaría del 17,7 al 46,5 %.

3.2.1.2. *En la interpretación de los resultados y en el uso del sistema*

Puede ocurrir también que el sesgo no se incorpore al diseñar el algoritmo o elegir los conjuntos de datos, sino a la hora de interpretar sus resultados o tomar decisiones en base a ellos que pueden tener efectos discriminatorios³²⁹. Por ejemplo, fijar el horario de trabajo de los empleados de una empresa en base a la organización previa llevada a cabo por el sistema de IA podría tener un impacto negativo en ciertos trabajadores, llegando incluso a ser discriminatorio.

Pensemos que las empresas, con el fin de gestionar sus recursos de la forma más eficiente posible, pueden utilizar sistemas de organización de horarios «*just-in-time*» para saber el número de trabajadores que necesitan en

conociendo el resultado o meta que se quiere alcanzar será posible introducir los datos necesarios para ello y establecer las instrucciones que se deberán seguir.

³²⁸ Este algoritmo sesgado, aplicado a casi 200 millones de estadounidenses cada año, se estima que redujo el número de pacientes negros identificados para recibir programas de atención médica a más de la mitad. Además, como advierten los especialistas, lo peor era que el mismo mecanismo sesgado se estaba usando con millones de personas en todo el mundo y en muchos otros ámbitos.

³²⁹ GERARDS, J./XENIDIS, R., *op.cit.*, pp. 86-87. En el Informe *Algorithms and Human Rights*, elaborado por el Comité de Expertos en intermediarios de internet del Consejo de Europa (MSI-NET), p. 28, se pone de relieve que el desafío no solo está en los datos o en los prejuicios de las personas, sino en la forma en que los seres humanos usan, perciben e interpretan los resultados de los sistemas y en la falsa creencia de que los algoritmos son neutrales e imparciales.

cada momento del día en función de la demanda. Esto no es, en principio, una práctica discriminatoria. No obstante, este tipo de programación ha sido objeto de críticas debido a sus posibles efectos negativos en los empleados, ya que puede ocurrir que en base a éste y otros datos introducidos en el sistema y con el objetivo de conseguir una mayor rentabilidad empresarial, el algoritmo diseñe horarios fragmentados e incompatibles con las normas sobre tiempo de descanso o conciliación personal y familiar. Además, estos horarios cambiantes e impredecibles podrían afectar a aquellas personas cuyas responsabilidades familiares harían muy difícil (o más bien imposible) trabajar en esas condiciones, dándose la circunstancia de que dichas personas suelen ser mayoritariamente las mujeres, quienes acabarían perdiendo oportunidades laborales en más casos que los hombres. Para evitar este efecto, el empresario no debería asumir acríticamente el resultado algorítmico sino intervenir para corregir el sesgo (aun a costa de la eficiencia empresarial y ahorro de costes), o incluso, ignorar la decisión del sistema si con ella no puede garantizar la igualdad y no discriminación en las condiciones laborales³³⁰.

Otro ejemplo de decisiones humanas discriminatorias basadas en resultados o decisiones algorítmicas podría ser el del servicio público de desempleo de Austria, cuyo algoritmo trataba de clasificar a las personas desempleadas en función de sus probabilidades de encontrar trabajo con el objetivo de llevar a cabo una asignación de recursos públicos lo más eficiente posible. Como consecuencia de los criterios o características personales valoradas por el algoritmo, se acababa clasificando a los sujetos en un grupo u otro en función de si su probabilidad (estimada) de encontrar un nuevo trabajo era *alta*, *media* o *baja*³³¹. El sistema reflejaba las prácticas discriminatorias tradicionalmente

³³⁰ GINÈS i FABRELLAS, A., *op.cit.*, p. 310, señala a Starbucks como una de las empresas que utilizan estos sistemas. No obstante, la propia empresa reconoció los desafíos asociados con esta práctica y tomó medidas para mejorar la estabilidad y previsibilidad de los horarios de sus empleados, con el objetivo de mitigar posibles efectos negativos y promover un entorno laboral más equitativo.

³³¹ Sobre este asunto puede verse también el trabajo de FRÖHLICH, W./SPIECKER, I./DÖHMANN, G., «Können Algorithmen diskriminieren? (¿Pueden los algoritmos discriminar?)», *Verfassungsblog*, 26 de diciembre de 2018, en <https://verfassungsblog.de/koennen-algorithmen-diskriminieren/>, traducida en <https://www.lamarea.com/2019/05/17/pueden-discriminar-los-algoritmos/>, 17 de mayo de 2019 (última visita el 8 de mayo de 2024). Según señalan estas autoras, «el software divide a las personas que buscan empleo en tres grupos: las que tienen buenas, medianas o malas expectativas de encontrar un trabajo. Sobre esta base, el Servicio Público de Empleo de Austria (AMS por sus siglas en alemán) quiere concentrar sus recursos a partir de 2020 principalmente en las personas que se encuentran en el grupo medio, donde serán utilizados de forma más efectiva. Las “posibilidades de integración en el mercado laboral” de las mujeres, según lo que evalúa el algoritmo, son totalmente negativas. Además, tener hijos dependientes provoca a una mala clasificación, pero solo para las mujeres. En el caso de los hombres, los programadores de software argumentan que la obligación de cuidado no tiene un impacto negativo estadístico en sus expectativas en el mercado laboral».

seguidas en el mercado laboral y otorgaba peores puntuaciones, por ejemplo, a mujeres, personas con discapacidad o extranjeros. En opinión de SORIANO ARNANZ, el sistema provocaba la estigmatización de estos grupos y de sus miembros al clasificar a las personas más vulnerables en la categoría inferior. El servicio público de desempleo austriaco defendió la necesidad de que el sistema funcionara así con el fin de poder garantizar que las personas con mayores dificultades de encontrar trabajo recibieran mejores ayudas que otras. No obstante, a pesar de dicha justificación, se priorizó la eficiencia en el reparto de recursos sobre cualquier otro objetivo y, puesto que el sistema concluyó que la asignación de recursos más eficaz se alcanzaba asignando mayores ayudas a las personas con una *probabilidad media* de acceder al mercado laboral, la Administración decidió reducir las ayudas a los desempleados a quienes el programa asignó el nivel más bajo en función de sus escasas probabilidades de encontrar trabajo (esto es, mujeres, discapacitados y mayores de 30 años), perpetuando así el trato desfavorable y la discriminación a ciertos grupos sociales históricamente oprimidos³³².

Por otro lado, debe tenerse en cuenta que en algunos casos el carácter discriminatorio y el efecto estigmatizante del sistema puede residir en el propio uso o aplicación incorrecta que se hace del mismo. Por ejemplo, en el caso del algoritmo SyRI, el Tribunal de la Haya estimó que el sesgo se encontraba *en el propio uso del sistema* en tanto en cuanto se utilizaba exclusivamente en barrios donde vivían personas con rentas bajas o pertenecientes a minorías, porque eran percibidas como zonas cuya población tenía mayor predisposición a defraudar a Hacienda o a la Seguridad Social, lo que reforzaba la estigmatización de estos colectivos, limitando cada vez más sus posibilidades de percibir beneficios sociales a los que tendrían derecho por su condición socioeconómica³³³.

Por último, los algoritmos que funcionan bien en un contexto social o geográfico concreto podrían discriminar si se extienden a otro diferente (por ejemplo, porque no fueran lo suficientemente representativos de la población en la que se pretenden aplicar). Por ello, hay que ser cautos a la hora de extrapolar algoritmos que han sido diseñados para un contexto determinado a otro contexto distinto, ya que el mismo algoritmo podría tener un impacto neto positivo cuando se aplica en un entorno determinado y un efecto muy diferente cuando se aplica a otro³³⁴.

³³² SORIANO ARNANZ, A., «Decisiones automatizadas y discriminación...», *cit.*, pp. 7-8/31.

³³³ En opinión del Relator especial sobre Pobreza Extrema y Derechos Humanos de la ONU, era muy llamativo que un país calificado en ocasiones de paraíso fiscal, no se utilizara esta herramienta para los grandes contribuyentes para proteger su derecho a la intimidad, y sin embargo se empleara en áreas con una alta proporción de residentes de bajos ingresos, migrantes y minorías étnicas.

³³⁴ Estas cautelas se recogen en el RIA.

3.2.2. Los conjuntos de datos

En tanto que los algoritmos funcionan con datos, dependerán en gran medida de ellos. Por ende, aunque el algoritmo esté bien diseñado, si la información que se le proporciona es incorrecta, anacrónica o poco representativa, contaminará los resultados pudiendo generar decisiones erróneas, parciales, sesgadas o discriminatorias.

La existencia de conjuntos de datos sesgados es un factor muy importante que explica por qué un algoritmo puede llegar a discriminar³³⁵. De hecho, el NIST lo destaca como el primer reto para abordar el sesgo de la IA y lo relaciona con factores como la disponibilidad, representatividad y los sesgos sociales incorporados.

Ya en el año 2017, el Parlamento Europeo en su *Resolución sobre las implicaciones de los macrodatos en los derechos fundamentales*, hacía hincapié en que, como consecuencia de los conjuntos de datos y sistemas de algoritmos que se utilizan al hacer evaluaciones y predicciones en las distintas fases del tratamiento de datos, los macrodatos podían conllevar violaciones de los derechos fundamentales de los individuos, tratamientos diferenciados y discriminación indirecta de grupos de personas con características similares, por ejemplo, a la hora de contratar o evaluar a las personas o al determinar los nuevos hábitos de consumo de los usuarios³³⁶.

No puede desconocerse que las bases de datos disponibles pueden contener errores o sesgos provenientes, por ejemplo, de decisiones discriminatorias anteriores o de una representación desigual de determinados grupos

³³⁵ De hecho, a juicio de la mayoría de expertos, es la principal causa que origina la discriminación algorítmica. Vid., entre otros muchos, los autores citados *supra*, nota 311.

³³⁶ *Resolución del Parlamento Europeo, de 14 de marzo de 2017, sobre las implicaciones de los macrodatos en los derechos fundamentales: privacidad, protección de datos, no discriminación, seguridad y aplicación de la ley* (2016/2225(INI)) (2018/C 263/10). En esta Resolución, se insta a la Comisión, a los Estados miembros y a las autoridades encargadas de la protección de datos a que definan y adopten las medidas que se impongan para minimizar la discriminación y el sesgo algorítmicos y a que desarrollen un marco ético común sólido para el tratamiento transparente de los datos personales y la toma de decisiones automatizada que sirva de guía para la utilización de los datos y la aplicación en curso del Derecho de la Unión; se pide que evalúen de manera específica la necesidad, no solo de transparencia algorítmica, sino también de transparencia en relación con posibles sesgos en los datos de capacitación utilizados para hacer inferencias sobre la base de los macrodatos; y se recomienda que las empresas lleven a cabo evaluaciones periódicas sobre la representatividad de los conjuntos de datos, que consideren si los conjuntos de datos se ven afectados por elementos sesgados, y que desarrollen estrategias para superarlos; pone de relieve la necesidad de examinar la exactitud e importancia de las predicciones basadas en el análisis de los datos teniendo presente las preocupaciones de carácter ético y la equidad (Consideraciones Generales 20, 21 y 22).

basada en discriminaciones históricas³³⁷. Como pone de relieve el NIST, en los conjuntos de datos pueden concurrir sesgos sistémicos y sesgos estadísticos y computacionales: los primeros –también llamados sesgos institucionales o históricos– son el resultado de procedimientos y prácticas institucionales que dan lugar a que ciertos grupos sociales se vean favorecidos y otros se vean perjudicados o devaluados simplemente porque tradicionalmente ha sido así³³⁸; los segundos, se producen cuando la muestra utilizada no es representativa de la población y surgen de errores sistemáticos (no aleatorios) sin necesidad de que concurra un ánimo de discriminar o prejuicios preexistentes.

Está demostrado que los algoritmos entrenados con datos sesgados o no suficientemente representativos arrojan resultados discriminatorios. Ejemplos de casos donde el origen de la discriminación se situó en el conjunto de datos de entrenamiento fueron los de diversas compañías que utilizaban sistemas de reconocimiento facial como Microsoft, IBM y Face ++. Tras una investigación llevada a cabo en el MIT, a cargo de JOY BUOLAMWINI, se demostró que fallaban en un porcentaje mucho más elevado con mujeres negras que con hombres blancos y que la causa se debía a que habían sido entrenado con una cantidad de imágenes mucho mayor de los segundos que de las primeras³³⁹. Es decir, en los datos de referencia había un grupo demográfico subrepresentado que acabó provocando discriminación. En concreto, el porcentaje de error de los sistemas de reconocimiento facial era de un 1% en el caso de hombres blancos y de un 35% en mujeres de piel oscura.

³³⁷ CASTILLO OCARANZA, C., *op.cit.*, p. 76, señala que los datos de entrenamiento pueden provenir de varias fuentes, entre las que menciona las «decisiones históricas» adoptadas por humanos (en cuyo caso se corre el riesgo de reproducir o exagerar los sesgos históricos de personas o instituciones) y las «fuentes expertas» que se utilizan por ejemplo en el caso de sistemas de apoyo a la decisión en ámbitos como la medicina, donde los datos de entrenamiento pueden consistir en descripciones de signos y síntomas asociados a diagnósticos clínicos realizados por un médico/a para un conjunto de pacientes (en cuyo caso los sesgos pueden aparecer durante el etiquetado o en la elección de los casos que son etiquetados). También señala los «contenidos e interacciones producidos por usuarios finales» como otro tipo de dato de entrenamiento que se utiliza (en estos casos, cada conjunto de datos tiene sus propios sesgos y por ello es importante documentar quién está involucrado en su construcción, cómo se procesaron y limpiaron los datos y con qué fin se recogieron).

³³⁸ Como apunta FERNÁNDEZ HERNÁNDEZ, C., «Modelos para identificar y gestionar los sesgos en la Inteligencia Artificial», *Diario La Ley*, N° 60, Sección Ciberderecho, 21 de Marzo de 2022, p. 4/8, esto no tiene por qué ser el resultado de prejuicios o discriminación consciente, sino de que la mayoría siga las reglas o normas existentes, se trata de sesgos presentes en la cultura y sociedad en general.

³³⁹ BUOLAMWINI, J. «How I'm fighting bias in algorithms», *TEDx Beacon Street*[Video File], 2016 y BUOLAMWINI, J./GEBRU, T. «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification», *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, en *Proceedings of Machine Learning Research*, vol. 81, p. 1-15. Disponible en: <https://proceedings.mlr.press/v81/buolamwini18a.html>

Se ha afirmado que los sistemas de reconocimiento biométrico, y más en concreto los de reconocimiento facial y de voz, son especialmente sensibles a la discriminación³⁴⁰. En ambos casos, los fallos de dichos sistemas discriminan a las mujeres, bien porque reconocen peor los rostros de mujeres de piel oscura, bien porque tienen problemas con las voces más agudas, lo que también afecta principalmente a las mujeres. Y también en ambos casos el origen de la discriminación reside en que los sistemas se basan en conjuntos de datos en que predominan los rostros y voces de hombres blancos.

Otro ejemplo de sesgo en los datos es el de aquellos sistemas de IA que se entrenan con bases de datos policiales que contienen mayoritariamente detenciones y condenas de personas pertenecientes a minorías étnicas y raciales como consecuencia de la discriminación sistémica sufrida históricamente por estos grupos en sus relaciones con la justicia y los cuerpos y fuerzas de seguridad. En estos casos, el algoritmo aprende que la pertenencia a una minoría étnica o racial está relacionada con un mayor índice de criminalidad y en base a ello puede recomendar, por ejemplo, aumentar la vigilancia policial en determinados barrios donde residen estas minorías o no conceder la libertad provisional a un preso por alta probabilidad de reincidencia durante el permiso³⁴¹. Un caso de este tipo fue el del sistema *Compas* anteriormente analizado.

Siendo esto así, con independencia del diseño del sistema, si los datos de entrenamiento reflejan discriminaciones sistémicas, racismo estructural o inequidad de políticas o prácticas judiciales, los resultados que arroje darán

³⁴⁰ Así lo afirma respecto a los de reconocimiento facial, FERNÁNDEZ DE LA MORENA, B., *Discriminación algorítmica. Estudio del sesgo en arquitecturas de aprendizaje profundo*, UAM, Madrid, 2019, p. 2, citando investigaciones que han mostrado una fuerte discriminación algorítmica sobre los grupos infra-representados en las bases de datos más populares de imágenes faciales. Sobre las aplicaciones de reconocimiento facial, en el *Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión* de la Asamblea General de la ONU (A/73/348, p.15/ 24), se dice que «son imperfectas porque se basan en conjuntos de datos en que predominan los hombres blancos y tienen errores de hasta 20% para las mujeres y las personas de piel oscura. Cuando esos sistemas se usan, por ejemplo, para clasificar imágenes obtenidas con un motor de búsqueda, su potencial de discriminación puede traducirse en injerencias concretas en el ejercicio de los derechos de la persona de buscar, recibir e impartir información y de reunirse o asociarse libremente». Respecto a los sistemas de reconocimiento tanto de imagen como de voz, puede verse el estudio de CASTAÑEDA, J. y otros, «Dealing with Gender Bias Issues in Data-Algorithmic Processes: A Social-Statistical Perspective», *Algorithms* 2022, 15, 303. <https://doi.org/10.3390/a15090303>. Recuperado de <https://www.uoc.edu/es/news/2022/297-algoritmos-genero>.

³⁴¹ Puede verse el ejemplo que explica ORWAT, C., *op.cit.*, pp. 45-46 sobre el sistema «PredPol». En relación a los sistemas de vigilancia policial explica EGUÍLUZ CASTAÑEIRA, J.U., *op.cit.*, p. 347, que los efectos de una muestra sesgada pueden amplificarse por las predicciones algorítmicas al convertirse éstas en «profecías autocumplidas». En consecuencia, un barrio no aumentará su índice de criminalidad porque haya más delitos, sino simplemente porque se aumentará la vigilancia policial en esa área en base a una recomendación sesgada del sistema.

lugar, nuevamente, a discriminaciones. En otras palabras, si el algoritmo contiene datos sesgados y aprende de ellos es lógico que perpetúe y refuerce los sesgos aprehendidos³⁴².

En todos los ejemplos anteriores lo que realmente sucede es que, como los modelos son entrenados de acuerdo a una proporción no representativa, los algoritmos acaban ajustando el modelo a favor de la clase mayoritaria, derivando en métricas de exactitud sesgadas. Es decir, los atributos y rasgos del grupo dominante se magnifican y se toman como estándar de referencia para evaluar y adoptar decisiones, con lo que el algoritmo sigue beneficiando al grupo dominante³⁴³.

En muchos casos es la imposibilidad de disponer de información variada lo que da lugar a modelos sesgados. Esto exige adoptar mecanismos de mitigación, esto es, equilibrar previamente y tratar tales datos antes de que sean a su vez objeto de tratamiento por parte del algoritmo en cuestión³⁴⁴.

Debe abogarse por el uso de conjuntos de datos de calidad, precisos y fiables para el etiquetado, retroalimentación y aprendizaje del sistema. Habrá que ser especialmente cautos a la hora de seleccionar los datos, definir cuáles serán relevantes y determinar los atributos a los que se dará mayor peso o relevancia para la toma de la decisión³⁴⁵.

³⁴² A este «bucle de retroalimentación» hace referencia el RIA en su Considerando 67: «Los conjuntos de datos deben tener las propiedades estadísticas adecuadas, también en lo que respecta a las personas o los colectivos de personas en relación con los que esté previsto utilizar el sistema de IA de alto riesgo, prestando una atención especial a la mitigación de los posibles sesgos en los conjuntos de datos que puedan afectar a la salud y la seguridad de las personas físicas, tener repercusiones negativas en los derechos fundamentales o dar lugar a algún tipo de discriminación prohibida por el Derecho de la Unión, especialmente cuando los datos de salida influyan en la información de entrada de futuras operaciones (bucles de retroalimentación)».

Sobre esta «retroalimentación» de los sistemas con decisiones sesgadas también puede verse, ADAMS-PRASSL, J. et al, *op.cit.*, p. 152. Para estos autores, en el contexto de la contratación, por ejemplo, si un algoritmo de calificación crediticia sesgado predice que un grupo concreto tiene mayores probabilidades de reembolsar sus préstamos, se concederá un mayor número de créditos a los miembros de ese grupo y éste estará cada vez más sobrerrepresentado en los datos de reembolso de préstamos.

³⁴³ FERNÁNDEZ DE LA MORENA, B., *op.cit.*, p.2; GINÈS i FABRELLAS, A., *op.cit.*, p. 307.

³⁴⁴ Esta idea la pone de manifiesto la Comisión Europea en su Comunicación titulada «Generar confianza en la inteligencia artificial centrada en el ser humano» (COM/2019/168 final), p.5, donde señala que los datos que se recopilan pueden reflejar sesgos sociales o contener inexactitudes o errores, y esto debe resolverse *antes de entrenar el sistema con dichos datos*. Afirma también que los procesos y conjuntos de datos utilizados deben ponerse a prueba y documentarse en cada fase, como la planificación, el entrenamiento, el ensayo y el despliegue.

³⁴⁵ Como apunta SORIANO ARNANZ, A., «Decisiones automatizadas y discriminación...», *cit.*, p. 6/31, la selección y peso otorgado a las variables empleadas por estos sistemas para medir y predecir el objetivo al que se dirigen puede condicionar y sesgar los resultados de estos programas. Así, si

La relación entre el buen funcionamiento de los sistemas de IA y los datos de entrenamiento ha sido subrayada por las instituciones europeas en numerosas ocasiones. Por ejemplo, la Comisión, en su *Comunicación titulada «Generar confianza en la inteligencia artificial centrada en el ser humano»*, aseveraba que la calidad e integridad de los datos utilizados era «primordial» para el correcto funcionamiento de los sistemas de IA y que debían establecerse requisitos para garantizarla³⁴⁶. Tiempo después, en el *Libro Blanco sobre la IA*, volvía a propugnar la necesidad de adoptar medidas para garantizar que los sistemas de IA se entrenasen con conjuntos de datos suficientemente amplios y que englobasen todos los escenarios pertinentes para evitar situaciones peligrosas³⁴⁷. Y como no podía ser de otra forma, el RIA incide en esta idea y exige como requisito de obligatorio cumplimiento para los sistemas de alto riesgo el uso de datos de calidad³⁴⁸.

Por su parte, el Parlamento Europeo en su *Resolución sobre los procesos automatizados de toma de decisiones: garantizar la protección de los consumidores y la libre circulación de bienes y servicios, de 12 de febrero de 2020 (2019/2915(RSP))*, destacaba la importancia de usar conjuntos de datos no sesgados y de calidad con el fin de mejorar los resultados de los sistemas algorítmicos y de reforzar la confianza y la aceptación de los consumidores³⁴⁹. Y a nivel interno, organismos públicos como la AEPD apuntan que la existencia de sesgos en los modelos de inferencia está íntimamente ligada con la exactitud o calidad del dato y que el grado de calidad de los datos de entrenamiento no se mide simplemente por la acumulación de datos, sino por los parámetros de relevancia, actualidad, fiabilidad, robustez y extensión de la tipología de datos a los ámbitos relevantes del tratamiento³⁵⁰.

En definitiva, antes de proceder al desarrollo de un sistema de IA debería determinarse si existen conjuntos de datos *adecuados* para alcanzar el propósi-

un sistema empleado por un banco para predecir el riesgo de impago de préstamos y decidir sobre su concesión o denegación, prioriza el nivel de ingresos sobre la capacidad de ahorro como elemento indicativo de la capacidad crediticia de una persona, esto perjudicará a las mujeres en mayor medida que a los hombres como consecuencia de la brecha salarial tradicionalmente existente por razón de género. Sin embargo, la capacidad de ahorro de las mujeres es mayor que la de los hombres.

³⁴⁶ Vid., p. 5.

³⁴⁷ Vid., p. 23.

³⁴⁸ Incidiremos en esto más adelante.

³⁴⁹ Vid., ap. 12. Y en el ap. 13 subraya la necesidad de que estos sistemas no solo utilicen conjuntos de datos no sesgados y de calidad, sino que también utilicen algoritmos explicables e imparciales. Además, considera necesario contar con estructuras de revisión dentro de los procesos de las empresas con el fin de subsanar posibles errores en las decisiones automatizadas y que los consumidores deben tener la posibilidad de solicitar la revisión y rectificación humanas de aquellas decisiones automatizadas que sean definitivas y permanentes.

³⁵⁰ Además, esta Agencia considera que en el caso del *machine learning* es preciso balancear la necesidad de datos para entrenar los sistemas y los riesgos para los derechos y libertades de los interesados.

to o tarea que se pretende conseguir, ya que, tanto sus predicciones como su propio aprendizaje dependerá de ellos. Dicha adecuación de los conjuntos de datos requiere prestar atención a tres factores: por un lado, a los métodos estadísticos para mitigar los problemas de representación; por otro lado, a los procesos para tener en cuenta el contexto sociotécnico en el que se está desplegando la aplicación; y por último, a la interacción de los factores humanos con el sistema durante todo su ciclo de vida³⁵¹.

3.2.3. Las correlaciones o inferencias derivadas del aprendizaje del sistema

Como ya se puso de relieve, las conclusiones a las que llegan estos sistemas de aprendizaje automatizado se basan en correlaciones que encuentran entre los datos que procesan. Correlaciones que, en ocasiones, pueden ser incorrectas o basarse en lógicas difíciles o imposibles de comprender y que son inesperadas, irracionales, irrelevantes, o incluso inaceptables para los humanos, cuyo razonamiento se basa en causalidades y en consideraciones normativas y éticas. Por ejemplo, un sistema de IA podría inferir que las personas que llevan calcetines negros tienen un mayor riesgo de sufrir infartos. Esto es una correlación a la que llega el algoritmo en base a la frecuencia con que esto ocurre en el conjunto de datos suministrado (las personas que llevan esos calcetines suelen sufrir infartos en un mayor porcentaje que otras) y no una relación de causalidad (llevar calcetines negros es lo que ocasiona infartos). También puede ocurrir que el propio sistema cometa errores a la hora de procesar los datos, por ejemplo, asociando unos datos a un determinado sexo.

El problema reside en que actualmente estos sistemas se consideran tan fiables que sus resultados, a pesar de basarse en correlaciones, se están tratando como si fueran relaciones causales aunque no exista prueba de la causalidad³⁵². Por si esto fuera poco, la situación se agrava porque la gran cantidad de datos procesados por los sistemas y la opacidad y complejidad técnica de algunos de ellos, impiden dar una explicación acerca de los motivos por los cuales el sistema arroja un resultado y no otro, es decir, impide descubrir en base a qué criterio o criterios el algoritmo acaba discriminando.

Es más, en los sistemas adaptativos que van aprendiendo mientras están funcionando y en los que el código cambia en función de los datos de entrada, ni siquiera va a ser posible prever ni anticipar en la fase de diseño que el

³⁵¹ FERNÁNDEZ HERNÁNDEZ, C., *op.cit.*, p. 6/8, sobre el trabajo del NIST.

³⁵² SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 91.

programa arrojará resultados sesgados porque los mismos provendrán de las correlaciones que va reconociendo el sistema en un gran conjunto de datos cambiante³⁵³. Al no estar predeterminada la lógica computacional, es importante que contengan un «*event data recorder*» ya desde el diseño³⁵⁴.

Por lo que se refiere a los sesgos discriminatorios, el problema acaece cuando el algoritmo tiene en cuenta características o variables protegidas (como la raza, el sexo o la discapacidad, entre otras) y las correlaciona directa o indirectamente con resultados negativos³⁵⁵. Por ejemplo, cuando ser mujer se correlaciona con un peor desempeño profesional, no debido a una relación causal, sino a que históricamente las mujeres han sido evaluadas peor que los hombres por hacer el mismo trabajo³⁵⁶. Lo llamativo del caso es que incluso cuando el sistema es entrenado para no tener en cuenta motivos protegidos, o cuando en los datos de entrada ni siquiera aparecen tales criterios, los algoritmos pueden encontrar otras variables o puntos de datos aparentemente no relacionados con la característica protegida, pero que de hecho sí lo están (los denominados «*proxies*»), que pueden llevar igualmente a resultados discriminatorios³⁵⁷.

Esto ocurrió en el caso del algoritmo que Amazon implementó para la selección de personal. Dicho sistema fue entrenado con datos históricos de la empresa (esto es, en base a los datos de las contrataciones de los últimos 10 años donde accedieron al puesto ofertado mayoritariamente hombres). El sistema, que no tenía acceso directo a los datos relacionados con el sexo de los candidatos, *aprendió a buscar e inferir la categoría protegida de otros datos* y comenzó a descartar currículums donde aparecían ciertas palabras de las que se podía inferir el género (en este caso femenino) del candidato (por ej. capitana del equipo de ajedrez, adjetivos generalmente utilizados por mujeres, currículums con nombres de universidades femeninas, etc.). A mayor abundamiento, esos puntos de

³⁵³ ALAMEDA CASTILLO, M.T., *op.cit.* p. 13 y *Libro Blanco sobre IA*. En tales casos, explica SOLAR CAYÓN, J.I., *op.cit.*, p. 159 «resulta patente que el control o validación de los resultados de los sistemas de Machine learning es esencial para identificar y mitigar aquellos sesgos originados como consecuencia del propio proceso de aprendizaje del algoritmo y que pueden resultar difíciles o imposibles de detectar en la fase de diseño.

³⁵⁴ Como explicaremos en su momento, esta medida se ha incorporado en el RIA.

³⁵⁵ GINÈS i FABRELLAS, A., *op.cit.*, p. 310.

³⁵⁶ Así se explica en GERARDS, J./ XENIDIS, R., *op.cit.*, p. 44. Sobre la diferencia entre correlación y causalidad y la repercusión que la misma tiene en este ámbito puede verse también, SÁNCHEZ CAPARRÓS, M., *op.cit.*, p. 5/34.

³⁵⁷ BAROCAS, S./SELBST, A. D., *op.cit.*, p. 691, mantienen que incluso en los casos en que se eliminan los criterios protegidos, los datos están libres de sesgos, las características son especialmente granulares y diversas, y el único objetivo es maximizar la precisión clasificatoria o predictiva, puede generarse discriminación debido a las «codificaciones redundantes», es decir, casos en los que la pertenencia a una clase protegida está codificada en otros datos.

datos no eran relevantes para predecir la variable objetivo, es decir, no tenían valor para predecir la idoneidad de los candidatos para el puesto de trabajo.

Tras varios intentos de modificar el comportamiento del sistema para conseguir mayor neutralidad, se acabó retirando porque no se tenía certeza de que el sesgo no pudiera reproducirse de nuevo por otra vía. No se llegó a identificar si el origen del sesgo estaba en los prejuicios de las personas que hasta ese momento se habían encargado de tomar las decisiones relativas a la contratación de personal, o si era el resultado de la menor presencia de mujeres en ámbitos tecnológicos, ya que esto contribuía a que la muestra de currículums de mujeres fuese menor. En cualquier caso, el sistema aprendió que ser mujer era equivalente a ser una mala candidata para puestos de trabajo en la empresa procediendo a eliminar dichos currículums³⁵⁸. Para evitar este resultado y mitigar el sesgo, hubiera sido conveniente dar al algoritmo indicaciones acerca de la igualdad de oportunidades que se buscaba.

Cuando se dice que la discriminación algorítmica puede deberse a los *proxies*, se está haciendo referencia a que los algoritmos toman decisiones basadas en variables que no son directamente discriminatorias, pero que están correlacionadas con características protegidas legalmente³⁵⁹. Como indica HUERGO LORA, en estos casos se alude a algo que está próximo o cerca de otra cosa a la que no tenemos acceso directo, pero que puede servir para aproximarnos a ella³⁶⁰.

El ejemplo mayoritariamente utilizado para explicar esta derivación de datos por proximidad con otros, es el del *código postal*. Éste dato puede usarse como *proxy* para determinar la nacionalidad, condición social, raza o etnia de una persona, así como su renta o capacidad económica³⁶¹. Aunque el código postal en sí mismo no es una característica protegida, puede estar correlacionado con ciertos grupos étnicos o niveles socioeconómicos debido a la segregación residencial.

³⁵⁸ En opinión de SORIANO ARNANZ, A./SIMÓ SOLER, E., *op.cit.*, p. 194 y ss, que compartimos, en el caso de Amazon la discriminación se debió a una combinación de causas o factores ya que las inferencias erróneas a las que llegó el algoritmo se basaron, a su vez, en los datos de entrenamiento que se le introdujeron.

³⁵⁹ Literalmente la palabra inglesa *proxy* significa apoderado, pero aquí ha de interpretarse como variable de sustitución o de proximidad.

³⁶⁰ HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, p. 45. Como ejemplo de lo afirmado, este autor hace referencia a los modelos que intentan detectar el virus causante de la Covid-19 sin realizar un test del virus. Dichos modelos buscan determinados rasgos que son más fáciles de localizar y que estadísticamente suelen aparecer en quienes están infectados y utilizan esos rasgos como *proxies*.

³⁶¹ ORWAT, C., *op.cit.*, p. 56. BATTAGLINI MANRIQUE DE LARA, M., «Cómo el Código Postal sirve para la discriminación de precios», a 18 de enero de 2020, <https://www.linkedin.com/pulse/c%C3%B3mo-el-c%C3%B3digo-postal-sirve-para-la-discriminaci%C3%B3n-de-manuela/>

Pensemos en un banco que utiliza un sistema de IA para predecir qué solicitantes de préstamos no podrán afrontar su devolución para, en base a ello, decidir su concesión o denegación. Tras analizar los datos históricos (esto es, los préstamos concedidos por la entidad en años anteriores) el sistema detecta que un alto porcentaje de los impagos fueron de personas que residían en una zona determinada. Automáticamente, el sistema concluye que las personas con un determinado código postal tienen más probabilidades de incumplir sus préstamos y, en consecuencia, les asigna un nivel de riesgo elevado que podrá llevar a la entidad (directamente o junto con otros criterios) a la denegación del crédito. Aunque en principio el código postal era un criterio aparentemente neutro para predecir la probabilidad de impago, se acaba convirtiendo en una variable *proxy* de la variable de impago. Si además resulta que en esa zona (o código postal) la mayoría de residentes pertenecen a una determinada raza, etnia o condición socioeconómica, nos encontramos con que ese dato (el código postal) refleja un criterio protegido (nacionalidad, raza, etnia...) y que la correlación a la que llega el sistema y a la que se otorga un peso o valor negativo, podría acabar discriminando indirectamente a ciertos colectivos³⁶². En definitiva, utilizar el código postal como criterio para acceder al crédito tendría un impacto negativo mucho mayor en miembros del grupo protegido que en los que no pertenecen a él, al ser muy superior el número de personas de ese grupo que residen en esa zona.

Otros ejemplos que pueden usarse como *proxy* para averiguar características protegidas de las personas son los «me gusta» en Facebook, que pueden revelar características personales de los usuarios como el sexo, orientación sexual, raza u opiniones políticas. A modo de ejemplo, la aplicación Magic Sauce creada por la Universidad de Cambridge, permite trazar tu perfil psicológico usando tu cuenta de Facebook. Utiliza un algoritmo que analiza el comportamiento de una persona en redes sociales para crear su perfil, haciendo predicciones sobre su personalidad, edad, nivel de inteligencia, orientación sexual, ideas políticas y religiosas o, incluso, nivel de satisfacción con la vida.

Como explica con detalle HUERGO LORA, jurídicamente el concepto de *proxy* es muy relevante si se tiene en cuenta que esos indicadores van a ser los

³⁶² BATTAGLINI MANRIQUE DE LARA, M., *op.cit.loc.cit.*, menciona otros casos en los que puede observarse este tipo de práctica y su efecto discriminatorio. Se trata de la discriminación de precios en el sitio web de la empresa Home Depot (una empresa estadounidense de mejora del hogar, bricolaje, etc). El minorista cobra precios más altos o bajos según el código postal de cada visitante individual, por lo tanto, se determina el precio en función de atributos personales, en este caso la ubicación. De la misma forma que en el caso del banco, si la ubicación o el código postal tiene una correlación con un grupo vulnerable, de nuevo nos encontraríamos ante un caso de *proxies*. Esta práctica es muy utilizada también por aerolíneas o empresas de alojamiento de viajeros. Así, la aplicación de reserva el mismo día, HotelTonight, introdujo dos características que muestran tarifas con descuento a los usuarios en función de su ubicación.

presupuestos de hecho en cuya virtud se va a tomar una decisión. En Derecho, las decisiones no se adoptan en función de cualquier indicador sino del presupuesto de hecho previsto en la norma. Por ende, cuando lo que se utiliza para decidir si se aplica o no la consecuencia jurídica concreta (ej. concesión de ayuda o subvención pública) no es el supuesto de hecho contenido en la norma sino otro hecho o circunstancia diferente correlacionada con él y que es más fácil detectar (*proxy*), se debe justificar adecuadamente ese cambio³⁶³. A lo que añade el autor, como segundo inconveniente del uso de *proxies* para la toma de decisiones algorítmicas, que partiendo de que hay circunstancias que no deben tenerse en cuenta para la aplicación de consecuencias jurídicas (por ejemplo, las del art. 14 CE), desconocer los *proxies* en base a los cuales el sistema de IA ha aplicado una consecuencia jurídica, supondría un límite al uso de algoritmos en la toma de decisiones por cuanto no podría excluirse que se hubieran utilizado *proxies* inaceptables, esto es, variables que reflejen características protegidas.

Los datos más inofensivos e insospechados pueden servir para predecir información sensible de la persona generando un efecto claramente discriminatorio. Precisamente por ello, los desarrolladores de algoritmos deben ser conscientes de este problema a fin de adoptar medidas para evitar que la toma de decisiones discriminatoria basada en *proxies* pase inadvertida.

Puede traerse a colación aquí el caso de la *start-up* Gild, donde el algoritmo desarrollado para ayudar a las empresas a encontrar candidatos para puestos de trabajo detectó que las personas que visitaban páginas web de manga japonés tenían mayores habilidades para programar. Nótese que, en principio, ese dato (actividad o presencia en páginas web, foros, etc.) no se considera un criterio o práctica discriminatoria. Sin embargo, en este caso se advirtió discriminación porque, como regla general, dichas páginas eran visitadas en mayor medida por hombres en su tiempo libre, lo que llevó al algoritmo a penalizar o descartar del proceso de selección laboral a las mujeres programadoras³⁶⁴.

³⁶³ HUERGO LORA, A., «Una aproximación a los algoritmos ...», *cit.*, p. 45.

³⁶⁴ Este caso lo recoge O'NEIL, *op.cit.*, pp. 120-121. Sobre el mismo explica GINÈS i FABRELLAS, A., *op.cit.*, p. 310, que la discriminación se debió a una combinación de causas: «la discriminación proviene de una correlación identificada por la base de datos, si bien probablemente impactada por la mayor presencia de hombres en la base de datos. Es decir, el algoritmo ha identificado una variable que se correlaciona con ser un excelente programador. Sin embargo, dada la composición mayoritariamente masculina de la industria tecnológica, esta correlación probablemente se refiera al grupo dominante, si bien se aplica de forma generalizada. En otras palabras, el algoritmo ha identificado qué variables se correlacionan con buenas habilidades de programación en los hombres, pero no tiene ni idea –dadas las diferencias de género existentes entre hombres y mujeres– de qué hace que una mujer programadora sea excelente. Puede que exista –o a lo mejor no– una característica que se correlaciona con buenas habilidades de programación entre las mujeres que se ha perdido entre los tributos del grupo dominante».

Capítulo III

Principales desafíos que plantea la discriminación algorítmica

1. INTRODUCCIÓN

Con carácter general, los retos que los sistemas de IA han generado para los juristas están directamente relacionados con sus características intrínsecas. Su complejidad técnica, interconectividad, opacidad, dependencia de datos, vulnerabilidad, apertura, autonomía y capacidad de autoaprendizaje, así como la pluralidad de agentes involucrados en su diseño, desarrollo, implementación y uso, dificultan la averiguación de *cuándo, cómo o por qué* han llegado a un determinado resultado y *quién debe responder* de los daños que se ocasionen, lo que sin duda justifica la necesidad de su estudio y ha llevado a las instituciones europeas a trabajar intensamente en estos temas desde sus primeras iniciativas en materia de IA.

A nuestro modo de ver, de todas las características mencionadas, es la *opacidad o falta de transparencia* la que se sitúa en el origen de la gran mayoría de problemas jurídicos que plantea la discriminación algorítmica y que tienen que ver con la dificultad para detectarla, probarla y determinar tanto su causa como el agente o agentes responsable/s de los daños que de la misma derivan³⁶⁵. Advertía el Parlamento Europeo en su *Resolución de 20 de octubre de*

³⁶⁵ Estos obstáculos son explicados por Orwat, entre otros autores, quien considera que detectar y probar la discriminación algorítmica es complicado debido a la opacidad de los sistemas, ya que la transparencia es limitada al estar protegidos los algoritmos como secretos comerciales, lo que dificulta que las personas afectadas identifiquen y demuestren los prejuicios o sesgos que les afectan. Vid., ORWAT, C., *op.cit.*, pp. 66 y ss. También en el estudio realizado por GERARDS, J./

2020, con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial (2020/2014(INL))³⁶⁶, que la opacidad algorítmica podía hacer extremadamente costoso, o incluso imposible, determinar *quién controlaba el riesgo asociado al sistema de IA*, o *qué código, entrada o datos provocaban en última instancia el funcionamiento lesivo*, o *la identificación de la relación de causalidad entre el daño y el comportamiento que lo causa*, con el resultado de que las víctimas podían quedar sin recibir una indemnización adecuada. Y la misma idea la reitera la Comisión a lo largo de su última *Propuesta de Directiva relativa a la adaptación de las normas de responsabilidad civil extracontractual a la inteligencia artificial*, de 28 de septiembre de 2022 (COM/2022/496 final)³⁶⁷ —en adelante, PDRC—.

Al estudio de estos tres problemas dedicaremos el presente capítulo, analizando con particular detalle el tema de la opacidad de los sistemas de IA y las medidas para prevenirla. El motivo de tratar estas cuestiones en este capítulo y no en el siguiente, dirigido concretamente al estudio de las medidas para prevenir daños derivados de la discriminación algorítmica, es que con los mecanismos tendentes a conseguir transparencia y explicabilidad no solo se contribuye a la prevención del daño sino también a conseguir su reparación, ya que cuanto más transparentes, explicables y accesibles sean los sistemas, mayor facilidad existirá para demostrar la existencia (o inexistencia) de sesgos o discriminación del algoritmo.

Como mantiene la Comisión Europea, cuando a pesar de las medidas preventivas o controles *ex ante* se sigan produciendo violaciones de los derechos fundamentales, todas las exigencias tendentes a garantizar la transparencia y trazabilidad de los sistemas, unidas a unos controles *ex post* sólidos que garanticen la intervención rápida de las autoridades públicas cuando surjan riesgos inesperados una vez puestos en circulación, permitirán (o eso se pretende) ofrecer a las personas afectadas una compensación efectiva³⁶⁸.

XENIDIS, R., *op.cit.*, p. 108, se destaca que los problemas de discriminación de los sistemas de IA pueden ser difíciles de detectar y confrontar debido principalmente a problemas de transparencia y responsabilidad.

³⁶⁶ Vid., Considerando H).

³⁶⁷ Vid., por ejemplo, pp. 1, 16 y Considerandos 3 o 27. En el Considerando 3 señala que «cuando la IA se interpone entre el acto u omisión de una persona y el daño, las características específicas de determinados sistemas de IA, como la opacidad, el comportamiento autónomo y la complejidad, pueden hacer excesivamente difícil, si no imposible, que el perjudicado satisfaga la carga de la prueba. En particular, puede resultar excesivamente difícil demostrar que un dato de entrada concreto del que es responsable la persona potencialmente responsable ha dado lugar a una información de salida específica de un sistema de IA que, a su vez, ha provocado el daño en cuestión».

³⁶⁸ Apartado 3.5 Exposición de Motivos de la propuesta inicial.

2. PRIMER OBSTÁCULO: DETECTAR LA DISCRIMINACIÓN

2.1. CONSIDERACIONES PREVIAS SOBRE LA OPACIDAD ALGORÍTMICA

La falta de transparencia tanto a la hora de *utilizar* estos sistemas para la toma de decisiones, como a la hora de *explicar* los motivos que las sustentan, hace que sea extraordinariamente difícil para el ciudadano averiguar que está siendo discriminado.

Para LAZCOZ MORATINOS la opacidad es la «preocupación que más atención ha despertado en la literatura al evaluar la aplicación de estos modelos algorítmicos en la toma de decisiones»³⁶⁹. Es por ello que, como se irá viendo a lo largo de estas páginas, en todos los instrumentos normativos vigentes o proyectados relacionados con la IA, tanto en el ámbito europeo como fuera de él, la transparencia se erige en pieza clave o principio rector a la hora de diseñar, desarrollar e implementar algoritmos para la toma de decisiones. De hecho, en las *Directrices éticas* del Grupo de Expertos en IA, se incluye como uno de los siete requisitos esenciales a cumplir por los sistemas para conseguir una IA fiable. Y en la *Resolución sobre aspectos éticos de la IA 2020*, se estableció que el marco regulador de la IA debía regirse por los principios de transparencia, explicabilidad, equidad, rendición de cuentas y responsabilidad, con el fin de potenciar las ventajas de los sistemas y proteger al mismo tiempo a los ciudadanos de los riesgos que implican. Dicha protección, tal como subraya el Parlamento, implica que los humanos puedan comprender las acciones de la IA, para lo cual es imprescindible que los desarrolladores e implementadores garanticen, dentro de los límites de lo técnicamente posible, que la tecnología se despliega y utiliza respetando los requisitos de transparencia y permitiendo la auditoría y la trazabilidad.

2.2. TRANSPARENCIA ALGORÍTMICA Y EXPLICABILIDAD

No es suficiente con exigir que la IA o los algoritmos sean transparentes o explicables, sino que habrá que determinar qué implican estas exigencias y cómo se controlará su cumplimiento³⁷⁰.

³⁶⁹ Así lo afirma LAZCOZ MORATINOS, G., «Análisis jurídico de la toma de decisiones algorítmica ...», *cit.*, p. 286. También TOLOSA, P./DIBO, C., *op.cit.*, p.187, consideran que los algoritmos deben ser auditables, transparentes, y explicables y si se logra implementar algoritmos con tales características seguramente las decisiones que se obtengan podrán ser más transparentes que las decisiones humanas bien intencionadas, pero inconscientemente sesgadas.

³⁷⁰ Para TOLOSA, P./DIBO, C., *op.cit.*, p. 188, es muy importante diseñar organismos que supervisen los sistemas. Consideran que, por muchos requisitos legales que se establezcan a la hora de

Según establece la *Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica*, el principio de transparencia algorítmica «consiste en que siempre ha de ser posible justificar cualquier decisión que se haya adoptado con ayuda de la IA y que pueda tener un impacto significativo sobre la vida de una o varias personas». Es decir, «siempre debe ser posible traducir los cálculos del sistema de inteligencia artificial a una forma comprensible para los humanos», matizándose en otros instrumentos de *soft law* que dicho principio «no consiste en revelar códigos, sino en hacer inteligibles los parámetros y criterios de las decisiones que se toman»³⁷¹. La finalidad principal de todo ello es que los usuarios de estos sistemas puedan entender su funcionamiento y resultados. En definitiva, podría decirse que la transparencia sería la capacidad o posibilidad de saber cuándo y por qué un sistema de aprendizaje automático ha funcionado bien o mal.

Según el Estudio de Impacto algorítmico del Ministerio de Interior de Países Bajos, la transparencia puede ser tanto *interna* como *externa*. La primera va dirigida principalmente a los sujetos de la cadena de valor de los sistemas (proveedores, responsables del despliegue, importadores, distribuidores, etc.) en beneficio de auditores, supervisores, autoridades o jueces, y es esencial para conocer y comprender su funcionamiento y detectar posibles

desarrollar algoritmos, si tras su implementación nos desentendemos de sus resultados e impacto, la discriminación podría pasar desapercibida: «supongamos que la regulación exige que los algoritmos para seleccionar personal sean de caja blanca y que para ser utilizados deba demostrarse que garantizan la neutralidad de género para ser implementados. Asumiendo que se autoriza a implementar un algoritmo de tales características en cierto mercado laboral, es razonable pensar que después de su implementación tendamos a despreocuparnos sobre sus resultados o su impacto y a confiar plenamente en el algoritmo, en la regulación y en la igualdad en las contrataciones. Sin embargo, por ejemplo, si las actualizaciones del sistema no se realizan adecuadamente o la actualización de datos no se controla, podría ocurrir que los resultados en el futuro no fueran los esperados al momento de su implementación y que esto pase completamente inadvertido. Así la regulación debe prever estos efectos y también debe prever mecanismos de medición de impacto y de revisión periódica si realmente se aspira que sea eficaz».

³⁷¹ Así, el Dictamen del CESE sobre *Inteligencia artificial: anticipar su impacto en el trabajo para garantizar una transición justa*; la Recomendación del Consejo OCDE sobre IA de 2019 también recoge entre sus principios la exigencia de que los sistemas de IA sean transparentes y explicables lo que implica proporcionar información significativa, adecuada al contexto y coherente; hacer comprensible los sistemas de IA, para que las partes interesadas sean conscientes de sus interacciones con estas herramientas, especialmente en el lugar de trabajo; permitir que los afectados por un sistema de inteligencia artificial entiendan su resultado y puedan cuestionarlo. Vid., p.5/7. Recuperado de: <file:///C:/Users/ASUS/Downloads/db5053b5-93e0-4cf5-a7cf-edce5ee6e893.pdf>. Según el Estándar de Transparencia Algorítmica desarrollado por la Oficina Central Digital y de Datos de Reino Unido, «la transparencia algorítmica significa dar a conocer cómo las herramientas algorítmicas apoyan las decisiones, incluyendo proporcionar información sobre dichas herramientas y las decisiones asistidas por algoritmos en un formato completo, abierto, comprensible, de fácil acceso y gratuito».

errores³⁷²; la segunda, va dirigida al público en general o al menos a todos los interesados o afectados por la IA y también se conoce como «explicabilidad pública»³⁷³.

La transparencia guarda una estrecha relación con el *principio de explicabilidad*. Así se reconoce, por ejemplo, en la *Recomendación de la UNESCO sobre la ética de la IA de noviembre de 2021*, donde se afirma que ambos conceptos están íntimamente relacionados entre sí, y con la responsabilidad, la rendición de cuentas y la confianza en los sistemas. No obstante, mientras la *transparencia* pretende proporcionar información para que las personas puedan comprender cómo se implementa cada etapa de un sistema de IA en función de su contexto, conocer los factores que influyen en una predicción o decisión específicas y si existen o no medidas de seguridad o de equidad (e incluso excepcionalmente, en casos de amenazas graves con repercusiones adversas para los derechos humanos, podría exigir que se compartiera el código fuente o los conjuntos de datos utilizados), la *explicabilidad* tiene como objetivo hacer inteligible la entrada, la salida y el funcionamiento de cada componente algorítmico y la forma en que contribuye a los resultados del sistema³⁷⁴.

Esta conexión entre la transparencia y la explicabilidad también la pone de relieve el Grupo de Expertos en sus *Directrices éticas para una IA fiable*, donde, además de consagrarse la explicabilidad como uno de los cuatro *principios éticos* que deben respetarse en el desarrollo, despliegue y utilización de los sistemas de IA para garantizar los derechos fundamentales, se incluye como *componente fundamental de la transparencia* junto con la trazabilidad y la comunicación³⁷⁵. No obstante, como aclara el Considerando 49, pese a que nume-

³⁷² A esta transparencia interna o comunicación de la información entre los distintos sujetos de la cadena de valor de los sistemas se le otorga una especial relevancia a lo largo del RIA, como después se explicará. También es a la que se refieren los organismos de normalización técnica NIST (EEUU), ISO, CEN CENELEC (UE).

³⁷³ Así lo explica COTINO HUESO, basándose en el Estudio citado. Vid., COTINO HUESO, L., «Transparencia de la inteligencia artificial pública: marco legal, desafíos y propuestas», *Actualidad Administrativa*, N° I, Sección Actualidad, Diciembre 2023. Recuperado de La Ley Digital, LA LEY 12767/2023. Para este autor, la transparencia y la explicabilidad permiten comprender y resolver problemas técnicos del funcionamiento del sistema, especialmente para comprender la cadena de causalidades.

³⁷⁴ Vid., ap.37 a 41.

³⁷⁵ *Directrices éticas*, p. 22. COTINO HUESO, L., «Transparencia de la inteligencia artificial pública...», cit., p. 3/23, califica el concepto de transparencia que proporciona el Grupo de expertos en IA de la Comisión Europea en sus *Directrices éticas* como un «concepto inclusivo». Nótese que dicho concepto de transparencia mantenido por el Grupo de Expertos en IA en sus *Directrices éticas*, ha sido el acogido en la versión final del RIA. Así, en el Considerando 27 se establece que, de acuerdo con las citadas Directrices, por transparencia se entiende que «los sistemas de IA se desarrollan y utilizan de un modo que permita una *trazabilidad y explicabilidad* adecuadas, y que, al mismo tiempo, *haga que las personas sean conscientes de que se comunican o interactúan con un sistema de IA e informe debidamente*

rosas obligaciones legales reflejan principios éticos, el cumplimiento de estos últimos trasciende el mero cumplimiento de las leyes existentes.

Como *principio rector*, se entiende que la explicabilidad «es crucial para conseguir que los usuarios confíen en los sistemas de IA y para mantener dicha confianza» e implica que «los procesos han de ser transparentes, que es preciso comunicar abiertamente las capacidades y la finalidad de los sistemas de IA y que las decisiones deben poder explicarse –en la medida de lo posible– a las partes que se vean afectadas por ellas de manera directa o indirecta», ya que, si no fuera así, se les estaría privando de la posibilidad de presentar alegaciones e impugnarlas adecuadamente³⁷⁶.

Como *elemento integrante de la transparencia*, se define como la «capacidad de explicar tanto los procesos técnicos de un sistema de IA, como las decisiones humanas asociadas» y se establece que dicha exigencia requiere que las decisiones del sistema «sean comprensibles para los seres humanos y estos tengan la posibilidad de rastrearlas»³⁷⁷.

Además de la explicabilidad, la transparencia incluye también la *trazabilidad* y la *comunicación*. Esto significa que no sólo se debe poder reconstruir cómo y por qué un sistema se comporta de determinada manera, sino que quienes interactúen con ellos deben saber que se trata de un sistema inteligente (y no de un humano) y qué personas serán sus responsables.

A mayor abundamiento, la transparencia también debe aplicarse a los elementos del sistema de IA, es decir, los datos, el *software* y los modelos de negocio³⁷⁸. Para conseguirla, deben documentarse los conjuntos de datos y procesos que dan lugar a la decisión del sistema, incluidos los relativos a la recopilación y etiquetado de datos, estar disponibles las explicaciones sobre el grado en que un sistema de IA condiciona e influye en el proceso de toma de decisiones de la organización, las opciones de diseño de dicho sistema y la justificación de su despliegue³⁷⁹. En relación a este último extremo, algunos

a los responsables del despliegue acerca de las capacidades y limitaciones de dicho sistema de IA y a las personas afectadas acerca de sus derechos». Este considerando 27 contiene lo dispuesto en la enmienda 213 del Parlamento europeo que proponía un nuevo art. 4 bis donde se recogieran los «Principios generales aplicables a todos los sistemas de IA», en cuyo apartado 1. d) figuraba la transparencia.

³⁷⁶ Vid., ap. 53 *Directrices éticas*.

³⁷⁷ *Directrices éticas*, p. 22.

³⁷⁸ *Ibidem*. También la Agencia Europea de los Derechos Fundamentales mantiene que, tanto la transparencia como las evaluaciones de impacto, deben verificarse tanto sobre la calidad de los datos, como sobre la forma en que funcionan los algoritmos (caja negra) y las facultades predictivas de los mismos. Vid., EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS (FRA), «BigData: Discrimination in data-supported decision making» 2018.

³⁷⁹ Vid., Comunicación de la Comisión Europea «Generar confianza en la inteligencia artificial centrada en el ser humano» (COM/2019/168 final), p. 6.

estudios advierten que puede haber otros problemas de transparencia que no estén relacionados con el algoritmo en sí mismo o con los datos que utiliza, sino con la falta de claridad en cuanto a las razones por las que se usa un determinado sistema de IA y los objetivos que persigue. En Francia, por ejemplo, el algoritmo *Parcoursup*, que se utilizaba para asignar a los estudiantes a las universidades, fue muy criticado porque sus objetivos no estaban bien definidos³⁸⁰.

Cuando los datos que empleen estos sistemas sean datos *personales* será de obligada aplicación el RGPD, en cuyo contexto la transparencia se configura como una obligación para los responsables del tratamiento a la hora de comunicar determinada información al interesado, lo que debe hacerse *ex art.* 12.1 de forma «concisa, transparente, inteligible y de fácil acceso, con un lenguaje claro y sencillo»³⁸¹. De ahí que se infiera que la principal preocupación del RGPD es la transparencia «hacia la persona interesada», más allá de las posibles soluciones técnicas para ayudar a esclarecer lo que hasta el momento es inescrutable³⁸².

Sin embargo, como apunta la AEPD, en los casos de *tratamientos de datos basados en IA* (esto es, cuando los sistemas de IA se incluyen en, o son medios de, un tratamiento de datos personales) «la transparencia debe permitir a los interesados ser conscientes del impacto que el empleo de dichas soluciones puede llevar asociado y de ahí que la transparencia esté dirigida tanto a los interesados como a los operadores del tratamiento, y que esté ligada, en particular, con una *información veraz sobre la eficiencia, las capacidades y las limitaciones reales de los sistemas de IA, que evite la creación de falsas expectativas, en los usuarios y los interesados, que puedan ocasionar una mala interpretación de las inferencias que se realizan en el marco del tratamiento. También, la transparencia está ligada a información sobre el contexto y situación del tratamiento, como la existencia de terceras partes, la localización física/virtual de la solución AI*, etc. En tales casos, los responsables del tratamiento deben obtener suficiente información sobre los sistemas para cumplir con las obligaciones que les impone el RGPD y que incluyen la transparencia para permitir el ejercicio de los derechos, el cumplimiento del principio de responsabilidad activa, o cumplir los requisitos de las Autoridades de Supervisión en relación con sus poderes de investigación y de los organismos de certificación y supervisión de códigos de conducta. Además, como sigue di-

³⁸⁰ GERARDS, J./ XENIDIS, *op.cit.*, p. 87.

³⁸¹ El principio de transparencia se consagra en el art. 5 RGPD y se desarrolla en los Considerandos 39 y 58, donde se establece que el mismo exige que toda información y comunicación relativa al tratamiento de los datos personales sea concisa, fácilmente accesible y fácil de entender, y que se utilice un lenguaje sencillo y claro, y, además, en su caso, se visualice.

³⁸² LAZCOZ MORATINOS, G., «Análisis jurídico de la toma de decisiones algorítmica...», *cit.*, p. 296.

ciendo la AEPD, «[l]a transparencia no se reduce a un instante puntual, sino que debe ser entendida como un principio en torno al que orbita de forma dinámica el tratamiento realizado y que afecta a todos y cada uno de los elementos y participantes que intervienen en la solución»³⁸³.

En consecuencia, es importante percatarse ya de entrada de que la transparencia tiene diferentes implicaciones en el RIA y en el RGPD en la medida en que afecta a distintos actores, se refiere a información diferente y va dirigida a diferentes destinatarios³⁸⁴. En el RIA la transparencia es exigible desde el diseño del sistema y a lo largo de todo su ciclo de vida, independientemente de si el mismo procesa o no datos personales. Va referida a los sistemas en sí mismos y las obligaciones que de ella derivan van dirigidas a diseñadores, desarrolladores, proveedores y responsables del despliegue. Por su parte, la transparencia recogida en el RGPD se aplica a los tratamientos de datos personales definidos en los arts. 2 y 4.2, y las obligaciones que implica recaen sobre los responsables de dicho tratamiento. No obstante, como aclara la AEPD, «los *diseñadores y desarrolladores* pueden ser responsables o encargados del tratamiento si utilizan datos personales en el diseño o desarrollo del sistema de IA. Los *proveedores* pueden ser responsables o encargados del tratamiento si los sistemas de IA almacenan o tratan datos de interesados identificados o identificables. [Y] los *usuarios/entidades que despliegan* un sistema de IA podrían ser responsables o encargados del tratamiento si incluyen dicho sistema como parte de sus tratamientos»³⁸⁵. En cualquier caso, sólo aquellos que actúen como responsables del tratamiento quedarán sujetos a las obligaciones de transparencia del RGPD.

Por último, cabe señalar también que, mientras que los destinatarios de las obligaciones de transparencia derivadas del RGPD son los interesados –personas físicas– y su contenido mínimo se recoge en los arts. 13 y 14 RGPD, la transparencia a la que se refiere el RIA se dirige principalmente a los responsables del despliegue y está relacionada con la explicabilidad, documenta-

³⁸³ Informe de la AEPD sobre la *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción*, 2020, p. 33.

³⁸⁴ Como explica la AEPD en <https://www.aepd.es/prensa-y-comunicacion/blog/inteligencia-artificial-transparencia>, el RIA (cuyo ámbito material son los sistemas de IA) establece un concepto de transparencia que difiere del mismo término establecido en el RGPD (cuyo ámbito material son los tratamientos de datos personales).

³⁸⁵ <https://www.aepd.es/prensa-y-comunicacion/blog/inteligencia-artificial-transparencia>. Sobre esto, el Considerando 10 RIA señala que el Reglamento no afectará a las obligaciones de los proveedores y responsables del despliegue en su papel de responsables o encargados del tratamiento si el diseño, desarrollo o uso del sistema IA implica tratamiento de datos personales. Las normas armonizadas del RIA deberán facilitar la aplicación efectiva y permitir el ejercicio de los derechos y otras vías de recurso de los interesados garantizados por el Derecho de la Unión en materia de protección de datos personales, así como de otros derechos fundamentales.

ción, mantenimiento de registros e información sobre cómo utilizar dicho sistema de IA. Tal como dispone el art. 13 del nuevo RIA titulado «*Transparencia y comunicación de información a los responsables del despliegue*», los sistemas de IA de alto riesgo «se diseñarán y desarrollarán de un modo que garantice que funcionan con un nivel de transparencia suficiente para que los *responsables del despliegue* interpreten y usen correctamente su información de salida [...]» (ap.1)³⁸⁶, e «irán acompañados de las instrucciones de uso correspondientes en un formato digital o de otro tipo adecuado, las cuales incluirán información concisa, completa, correcta y clara que sea pertinente, accesible y comprensible para los responsables del despliegue» (ap.2), indicando el tercer apartado el contenido de dicha información sobre el sistema, que expondremos en su momento. También en el Considerando 72 RIA se da amplia cuenta de que la transparencia exigible a los sistemas de IA de alto riesgo va dirigida, principalmente, a que los responsables del despliegue puedan «*comprender la manera en que el sistema de IA funciona, evaluar su funcionalidad y comprender sus fortalezas y limitaciones*», así como «*utilizar el sistema y tomar decisiones con conocimiento de causa*». No obstante, no cabe duda de que esto repercutirá indirecta y favorablemente en los usuarios finales³⁸⁷.

Además, como se explicará en su momento, aunque el RIA abogue principalmente por la transparencia interna, también establece obligaciones a favor de la transparencia externa y ha acabado incorporando en el art. 86, como novedad relevante respecto a su versión inicial, el *derecho a una explicación* que tendrán las personas afectadas por las decisiones algorítmicas³⁸⁸.

³⁸⁶ Sigue señalando el precepto que «se garantizará un tipo y un nivel de transparencia adecuados para que el proveedor y el responsable del despliegue cumplan las obligaciones oportunas previstas en la sección 3». Para determinar cuál es el nivel de transparencia adecuado debe tenerse en cuenta el equilibrio que la norma pretende conseguir entre la transparencia y explicabilidad y otros principios como la privacidad o seguridad. En cualquier caso, la UNESCO en su Recomendación ética IA 2021 señala que dicho grado de transparencia deberá ajustarse al contexto y efecto del sistema de IA.

³⁸⁷ De hecho, el art. 13.3.b) establece que las instrucciones de uso que deben suministrarse al responsable del despliegue deben contener, entre otros extremos, información sobre las características, capacidades y limitaciones del funcionamiento del sistema de IA de alto riesgo, y en particular, *sus capacidades y características técnicas para proporcionar información pertinente para explicar su información de salida* (art. 13.3.b.iv). Este último inciso parece contener, de forma algo sutil, la obligación del responsable del despliegue de explicar al afectado la decisión del sistema. Por tanto, el objetivo de la transparencia y comunicación de información a los responsables del despliegue no es solo que ellos mismos puedan comprender y utilizar adecuadamente el sistema de IA sabiendo, en general, cómo funciona y qué datos trata, sino también, en nuestra opinión, que sean capaces de explicar sus decisiones a los posibles afectados.

³⁸⁸ La transparencia hacia los usuarios finales también figura en el art. 50 (art. 52 en las versiones anteriores de la propuesta) y está relacionada con la obligación de advertir a las personas físicas de que están interactuando con un sistema de IA. Se trata de un precepto extenso que da contenido

2.3. OPACIDAD SOBRE EL USO DE SISTEMAS ALGORÍTMICOS DE TOMA DE DECISIONES Y SOBRE EL CONTENIDO DE LOS ALGORITMOS

Una vez destacada la relevancia y alcance que adquiere la transparencia en este ámbito, y antes de exponer qué medidas se están adoptando para garantizarla, conviene precisar que la opacidad algorítmica suele manifestarse tanto a la hora de proporcionar información suficiente sobre los sistemas que permita a los interesados saber cómo y en base a qué motivos se ha adoptado la decisión que les afecta, como incluso en relación al propio uso del sistema.

El Parlamento Europeo, en su *Resolución sobre aspectos éticos de la IA 2020*, hacía una sutil referencia a esta idea cuando distinguía entre la *transparencia de los algoritmos* y la transparencia en el *uso de los algoritmos*³⁸⁹. Es decir, en muchas ocasiones las personas ignoramos que estamos siendo objeto de decisiones algorítmicas y que las mismas nos discriminan injustamente³⁹⁰. Así, una persona que ha sido discriminada en una plataforma de Internet en relación con el precio de determinados productos en función de su perfil (incluido su género, edad, ubicación, etc.) puede que ni siquiera sepa que la oferta que

al capítulo IV de la Propuesta. Concretamente dispone, por un lado, que los proveedores deben garantizar que los sistemas de IA destinados a interactuar directamente con personas físicas se diseñen y desarrollen de forma que las personas físicas sepan que están interactuando con una IA excepto cuando resulte evidente desde el punto de vista de una persona física razonablemente informada, atenta y perspicaz, teniendo en cuenta las circunstancias y el contexto de utilización; además, cuando esos sistemas de IA (incluidos los de uso general) generen contenido sintético de audio, imagen, vídeo o texto, deben velar por que la información de salida esté marcada en un formato legible por máquina y que sea posible detectar que ha sido generada o manipulada de manera artificial. Por otro lado, respecto a los responsables del despliegue establece, en primer lugar, que cuando utilicen sistemas de reconocimiento de emociones o categorización biométrica deberán informar de su funcionamiento a las personas físicas expuestas a él, cuyos datos personales tratarán de conformidad con el RGPD y la Directiva 2016/680 (art.50.3), y en segundo lugar, que cuando utilicen sistemas que generen o manipulen imágenes o contenidos de audio o vídeo que constituyan una ultrafalsificación, deberán hacer público que estos contenidos o imágenes han sido generados o manipulados de manera artificial, y cuando se trate de texto que se publique para informar sobre asuntos de interés público deberán divulgar que el texto se ha generado o manipulado de manera artificial (art. 50.4).

³⁸⁹ Vid., ap. 20.

³⁹⁰ Expone esta situación SORIANO ARNANZ, A., «Decisiones automatizadas y discriminación...», *cit.*, p. 17/31, para quien, el acceso a la justicia para defender el derecho a la igualdad de trato y no discriminación no sólo se ve condicionado por cuestiones económicas o burocráticas, sino también por el hecho de que las víctimas de las situaciones de discriminación en muchas ocasiones no son tan siquiera conscientes de que sus derechos están siendo vulnerados. Aquellos que hayan sido discriminados injustamente en un proceso de selección laboral o en la concesión de un préstamo a través de sistemas automatizados que basan sus decisiones en motivos protegidos, difícilmente serán conscientes de las razones verdaderas por las que no se les contrata o se les deniega un préstamo, especialmente si no tienen acceso o conocimiento de otras personas que hayan participado en el mismo proceso.

se le muestra ha sido decidida por un sistema inteligente y menos aún que el mismo puede estar vulnerando sus derechos, sobre todo cuando no tiene una referencia comparativa, es decir, si no sabe que a otras personas que cuentan con atributos diferentes se le ofrecen los mismos productos a un mejor precio o condiciones³⁹¹.

También en su *Resolución de 12 de febrero de 2020, sobre los procesos automatizados de toma de decisiones*, advertía de que su uso planteaba retos para los consumidores que podían afectar tanto a su capacidad para *detectar* dichos procesos, como de *entender* su funcionamiento, *tomar decisiones con conocimiento de causa* en cuanto a su utilización y *autoexcluirse*³⁹².

En virtud de lo anterior, consideramos junto a un sector doctrinal que habría que distinguir dos niveles de opacidad algorítmica³⁹³. El primero sería aquel en que el interesado desconoce que es objeto de una decisión automatizada, y el segundo, aquel en que la persona sabe que la decisión está basada en un tratamiento automatizado de sus datos, pero no se le permite conocer el funcionamiento del proceso decisorio, impidiéndole conocer los motivos de la decisión que le afecta.

Como es lógico, si ni siquiera sabemos que estamos siendo objeto de decisiones algorítmicas, menos aún seremos capaces de descubrir que las mismas son discriminatorias y, en consecuencia, perderemos la oportunidad de impugnarlas en defensa de nuestros derechos³⁹⁴. De lo que no cabe duda es de que, en tales casos, la frustración de las legítimas expectativas de las personas

³⁹¹ Advierte BATTAGLINI MANRIQUE DE LARA, M., «Cómo el Código Postal sirve para la discriminación de precios», *cit.*, que es mucho más probable que la discriminación de precios se dé en mercados donde los consumidores no tienen información directa sobre cuánto pagan otros consumidores y en mercados donde los precios diferenciales se combinan con productos heterogéneos, de modo que las comparaciones entre consumidores sean difíciles. En el caso de la tarjeta *Apple Card*, pudo descubrirse la diferencia de tratamiento entre las condiciones ofrecidas a un hombre y una mujer en cuanto al límite de crédito de su tarjeta, precisamente porque estaban casados.

³⁹² Vid., ap. C).

³⁹³ LAZCOZ MORATINOS, G., *op.cit.*, p. 293; VESTRI, G., «La inteligencia artificial ante el desafío de la transparencia algorítmica. Una aproximación desde la perspectiva jurídico-administrativa», *Revista Aragonesa de Administración Pública*, n° 56, 2021, p. 382.

³⁹⁴ En la misma línea considera LLORENS ESPADA, J., «Responsabilidades civiles por discriminación por razón de género cuando medie un sistema de inteligencia artificial», en Rivas Vallejo (dir.), *Discriminación algorítmica en el ámbito laboral*, Navarra, 2022, p. 576, que en el ámbito laboral será difícil para un trabajador saber si una decisión empresarial esconde un proceso algorítmico, o incluso sabiéndolo, no será fácil obtener información de los motivos en base a los cuales el algoritmo toma las decisiones y de qué parte del proceso deriva la falta de equidad. Por ende, el trabajador no podrá saber que está siendo discriminado ni podrá incoar un procedimiento de tutela de sus derechos fundamentales ante casos de discriminación generados por algoritmos. En sentido parecido, EGUÍLUZ CASTAÑEIRA, J.U., *op.cit.*, p. 337.

o la pérdida de oportunidad de acceder a un empleo, conseguir un crédito, ingresar en un prestigioso centro educativo o ser beneficiario de un tratamiento médico o de un seguro de vida (entre otros muchos ejemplos que podrían ponerse) basadas en criterios injustos o discriminatorios, generan daños a personas concretas que en muchos casos quedarán sin resarcir por desconocimiento o ignorancia acerca de que, tras esas decisiones, se esconde un proceso algorítmico y de que el mismo arroja resultados discriminatorios.

Es más, aun siendo conscientes de que detrás de la decisión que nos afecta se esconde un proceso algorítmico, seguiría siendo muy difícil *detectar* posibles errores o casos específicos de discriminación teniendo en cuenta que la decisión algorítmica se basa en una combinación de múltiples variables con ponderaciones diversas a las que el ciudadano en principio no tiene acceso o, incluso teniéndolo, no comprende. Y es que en muchas ocasiones estos sistemas de IA son técnicamente tan complejos que ni siquiera los programadores pueden explicar las razones por las que ha llegado a una decisión y no a otra distinta.

Por ello se ha mantenido, con acierto a nuestro juicio, que uno de los rasgos definitorios de la discriminación algorítmica es precisamente su *invisibilidad o facilidad de pasar desapercibida*³⁹⁵. Urge por tanto que la ciudadanía tome plena conciencia de que hoy en día muchas de las decisiones que nos afectan están siendo adoptadas por algoritmos, o en base a sus resultados, que pueden contener sesgos que vulneran nuestro derecho a la igualdad de trato y a la no discriminación.

2.4. OPACIDAD JURÍDICA Y TÉCNICA DEL ALGORITMO

En el tratamiento del problema de la opacidad algorítmica inciden factores tanto técnicos como jurídicos. Es decir, la falta de transparencia puede

³⁹⁵ Así lo mantienen, entre otros, SÁEZ LARA, C., *op.cit.*, p. 45, para quien este rasgo es, a su vez, el principal problema de estas tecnologías; EGUÍLUZ CASTAÑEIRA, J.U., *op.cit.*, p. 337; y NAVAS NAVARRO, S., *op.cit.*, pp. 50-51. Esta última autora habla del *weblining* (tradicionalmente *redlining*) como forma invisible de discriminación y explica que consiste en segmentar a la población para denegarle la prestación de servicios u ofrecerle productos a precios superiores en función de su lugar de residencia sobre la base del código postal, lo que indicaba la pertenencia a una determinada raza o etnia. En estos casos, la discriminación deriva del comportamiento *online* del usuario de Internet, con la particularidad de que éste no es consciente de ello. Por tanto, es una discriminación invisible para el usuario llevada a cabo a través de programas de ordenador. En opinión de esta autora, que compartimos, la elaboración automatizada de perfiles hace prácticamente imposible detectar estas discriminaciones porque el usuario no sabe que se le está segmentando y, si lo es, no alcanza a comprender el significado que ello representa en relación con la formación de su voluntad, de su libertad y autonomía personal. Por tanto, no es sólo que haya asimetría en la información entre las partes sino que esta ocasiona discriminación.

deberse tanto a la propia complejidad intrínseca del algoritmo, cuya comprensión exige una capacidad analítica y unos conocimientos técnicos considerables, como al régimen jurídico que lo protege en tanto en cuanto pueda ser objeto de derechos de propiedad intelectual o secreto empresarial³⁹⁶.

2.4.1. La opacidad jurídica

2.4.1.1. Sobre el código fuente

El código fuente de un algoritmo recoge las instrucciones del programador para resolver un problema o realizar una tarea e incluyen, entre otros elementos, operaciones matemáticas, lógica, estructuras de control o funciones. Podría definirse como la representación escrita y legible por humanos de las instrucciones detalladas para resolver un problema específico utilizando un lenguaje de programación particular³⁹⁷. El objetivo del código fuente es crear normas y disposiciones claras para el ordenador y que éste sea capaz de tra-

³⁹⁶ La doctrina ha hecho referencia a esta idea de varias maneras. Por ejemplo, para LAZCOZ MORATINOS, G., *op.cit.*, p. 293, en el ámbito de la IA la opacidad es polisémica y puede utilizarse con un doble sentido: por un lado, como equivalente a *ininteligible* (inherente al *machine learning in a big data context*), y por otro lado, para designar la opacidad intencionada en base a razones de derechos de autor, secreto industrial, privacidad o seguridad. Para SUBERBIOLA GARBIZU, I., «Inteligencia artificial, opacidad y motivación de las decisiones administrativas automatizadas», en *La protección de los derechos fundamentales en el ámbito tributario*, coord. por A.Vázquez del Rey Villanueva e I. Suberbiola Garbizu, 2021, recuperado de La Ley Digital (LA LEY 1117/2021), hay que ser conscientes de que existen *black boxes* por la arquitectura del propio sistema de IA y *black boxes* inducidas por la normativa que lo regula. Y CASTILLO OCARANZA, C., «Discriminación algorítmica en el ámbito laboral», en P. Rivas Vallejo (dir.), *Discriminación algorítmica en el ámbito laboral*, Navarra 2022, p. 75, considera que la opacidad algorítmica podría ser incluso triple, ya que, junto a la opacidad legal y a la opacidad técnica debida a la complejidad inherente al propio sistema, habría que añadir la inducida a través de «mecanismos de ofuscación intencional» utilizados para alterar las instrucciones de un programa (esto es, su código fuente) de forma que, a pesar de seguir manteniendo su función original, se hicieran difíciles de leer.

³⁹⁷ El código fuente es la versión del *software*, ya que está *originalmente escrito* por un ser humano en texto plano (es decir, legible por humanos en caracteres alfanuméricos) y después se escribe en un lenguaje de programación determinado elegido por el programador que puede ser, entre otros, C++, Python, Java, etc, que sigue reglas y sintaxis específicas sólo descifrables por expertos en informática. Por tanto, el código fuente no solo debe comprenderlo el ordenador, sino que también debe ser comprensible por los programadores para facilitar su mantenimiento y posibles modificaciones en el futuro. Vid., sobre el concepto de código fuente, RIVAS VALLEJO, P. «Análisis...», *cit.*, p. 440 y fuentes citadas. También DE LA CUEVA, J., «Código fuente, algoritmos y fuentes del Derecho», *El Notario del siglo XXI*, n°77, enero-febrero 2018, pp. 36-39, accesible en <https://www.elnotario.es/index.php/hemeroteca/revista-77/opinion/opinion/8382-codigo-fuente-algoritmos-y-fuentes-del-derecho> (última consulta 14 de mayo de 2024).

ducirlas a su lenguaje. Así, el código fuente sería la base de los programas de ordenador y de las páginas web.

Ante la petición que un posible afectado por una decisión algorítmica basada en sus datos personales pudiera formular para conocer dicha información y tratar así de descubrir cómo se ha tomado la decisión que le afecta, la doctrina mantiene que podrían alegarse varios fundamentos jurídicos que blindarían el acceso:

En primer lugar, aquellos que consideran que la definición de «programa de ordenador» del art. 96.1 LPI es aplicable al algoritmo en tanto que secuencia de instrucciones que constituyen el motor de funcionamiento de cualquier programa informático, mantienen que quedaría protegido por los derechos de autor regulados por el título VII LPI³⁹⁸. Teniendo en cuenta que el Considerando 63 RGPD establece que el derecho de protección de datos «no debe afectar negativamente a los derechos y libertades de terceros, incluidos los secretos comerciales o la propiedad intelectual y, en particular, los derechos de propiedad intelectual que protegen programas informáticos», la norma permitiría ocultar el código fuente a las personas interesadas que interponen una reclamación en base a dicho Reglamento. Por tanto, quedaría amparado el derecho de propiedad intelectual o secreto comercial que afectase al algoritmo y a su código fuente, si estos pudieran quedar perjudicados por su exposición, siquiera privada, a la persona reclamante afectada por su aplicación³⁹⁹.

En segundo lugar, el Considerando 11 de la *Directiva 2009/24/CE sobre la protección jurídica de los programas de ordenador*, dispone que «debe establecerse claramente que solo se protege la expresión del programa de ordenador y que *las ideas y principios implícitos en los elementos del programa, incluidas las de sus interfaces*, no pueden acogerse a la protección de los derechos de autor con

³⁹⁸ RIVAS VALLEJO, P. «Análisis...», *cit.*, p. 442.

³⁹⁹ En este sentido RIVAS VALLEJO, P. «Análisis...», *cit.*, p. 439, para quien, en la medida en que el algoritmo es susceptible de titularidad o propiedad por parte de personas físicas dentro del contexto de los derechos de propiedad intelectual, el empresario podría negarse a explicar los extremos anteriores. Sin embargo, otro sector doctrinal defiende la prevalencia de la protección de datos personales sobre los secretos comerciales y los derechos de propiedad intelectual en base a los Considerandos 34 y 35 de la propia *Directiva 2016/943 sobre secretos comerciales*, que disponen que la misma respeta el derecho a la protección de datos personales al mismo tiempo que se respeta el secreto comercial, y que dicha Directiva no debe afectar a los derechos y obligaciones previstos en la normativa de Protección de Datos y en particular a los derechos del interesado de acceder a aquellos de sus datos personales que sean objeto de tratamiento. En base a ello, el derecho a la protección de datos del interesado debe prevalecer. Así lo afirman también las *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento General de Protección de Datos* al afirmar que los responsables no pueden basarse en la protección de sus derechos comerciales como excusa para denegar el acceso o negarse a ofrecer información al interesado. Vid., p. 19.

arreglo a la presente Directiva», disposición que después se recoge en su art. 1.2⁴⁰⁰ y que también establece el art. 96.4 LPI. De acuerdo con este principio de derechos de autor, sigue diciendo el Considerando 11, «*en la medida en que la lógica, los algoritmos y los lenguajes de programación abarquen ideas y principios, estos últimos no están protegidos con arreglo a la presente Directiva*». Precisamente por ello, algún autor mantiene que esta declaración excluiría la protección de los algoritmos por la vía de los derechos de autor⁴⁰¹. No obstante, el TJUE, en su sentencia de 22 de diciembre de 2010 (asunto *Bezpečnostní softwarová asociace*) entendió que el algoritmo constituye una creación intelectual, la que se documenta y escribe en lenguaje de código. Por tanto, como razona dicha sentencia, la Directiva protegería «el programa de ordenador en todas sus formas de expresión, que permiten reproducirlo en diferentes lenguajes informáticos, tales como el código fuente y el código objeto» (apartado 35). Además, ampararía una protección indirecta del algoritmo, cuando pudiera plasmarse por escrito, ya que la norma extiende la protección a toda la documentación técnica y los manuales de uso de un programa, donde puede recogerse el citado algoritmo (STJUE de 2 de mayo de 2012, asunto *SAS Institute Inc. v World Programming Ltd*), previsión que también contempla el art. 96.1 *in fine* LPI⁴⁰².

En tercer lugar, se ha mantenido que los algoritmos podrían quedar protegidos por la *Ley 1/2019, de 20 de febrero de 2019, de secretos empresariales* (en adelante, LSE), siempre que reúnan las condiciones necesarias para ser considerados como secreto empresarial, esto es, siempre que se trate de información no divulgada por las empresas, que tuviera valor comercial y se hubieran adoptado medidas razonables para su protección⁴⁰³.

⁴⁰⁰ Art. 1.2: «La protección prevista en la presente Directiva se aplicará a cualquier forma de expresión de un programa de ordenador. Las ideas y principios en los que se base cualquiera de los elementos de un programa de ordenador, incluidos los que sirven de fundamento a sus interfaces, no estarán protegidos mediante derechos de autor con arreglo a la presente Directiva».

⁴⁰¹ En este último sentido, ALAMEDA CASTILLO, M.T., *op.cit.* p. 34 o CASTILLO PARRILLA, J.A., «Sentencia del Tribunal Ordinario de Bolonia de 31 de diciembre de 2020 (Caso Deliveroo) ¿Discriminación algorítmica o discriminación a través de un algoritmo?», *Derecho Digital e Innovación. Digital Law and Innovation Review*, n.7 (octubre-diciembre), 2020, para quien la protección jurídica de los algoritmos se reduce a su protección como secretos comerciales, ya que no pueden ser objeto de derechos de autor ni son patentables.

⁴⁰² Art. 96.1 LPI 1. «A los efectos de la presente Ley se entenderá por programa de ordenador toda secuencia de instrucciones o indicaciones destinadas a ser utilizadas, directa o indirectamente, en un sistema informático para realizar una función o una tarea o para obtener un resultado determinado, cualquiera que fuere su forma de expresión y fijación. A los mismos efectos, la expresión *programas de ordenador* comprenderá también su documentación preparatoria. La documentación técnica y los manuales de uso de un programa gozarán de la misma protección que este Título dispensa a los programas de ordenador».

⁴⁰³ El TS en su sentencia n. 508/2018, de 20 de septiembre, confirmó que el código fuente de un programa informático podía estar protegido como secreto comercial si su divulgación suponía un

A la vista de estos argumentos, las empresas podrían ampararse en derechos de propiedad intelectual o secreto comercial para no ofrecer información acerca de los algoritmos en los que basan sus decisiones (aun cuando estos fueran explicables o de caja blanca) impidiendo al particular afectado investigar las razones por las que se ha llegado a la decisión que le perjudica e impugnarla cuando no comulgue con ella.

En cualquier caso, la protección ofrecida tanto por los derechos de propiedad intelectual como por el secreto empresarial se concede a los *autores*, según disponen el art. 96.2 LPI y el art. 1.2 de la LSE. Esto implica que, si la empresa que utiliza el sistema de decisión automatizada ha sido la creadora del sistema, quedaría protegida por dichas normas y podría oponerse a una eventual petición de acceso al código fuente del afectado por la decisión. Sin embargo, si como suele ocurrir, el sistema hubiera sido desarrollado por un tercero (por ejemplo, un proveedor tecnológico), el empresario que lo utiliza no podría oponer la propiedad intelectual para no revelar el código fuente ya que, sencillamente, no sería el propietario del mismo sino que, en virtud del art. 97 LPI, lo serían sus creadores o la persona jurídica a la que la ley reconociese la titularidad, incluyendo a las empresas para las que presten sus servicios los trabajadores que han diseñado el programa cuando se haya hecho en el marco de su relación de trabajo⁴⁰⁴.

En este último caso, esto es, cuando el sistema hubiera sido diseñado por un tercero y no por el propio empresario que lo utiliza (al que el RIA denomina finalmente «responsable del despliegue»⁴⁰⁵) habría que tener en cuenta dos cuestiones importantes ante una eventual solicitud de acceso al código fuente o información sobre el algoritmo por quien considera vulnerados sus derechos:

Por un lado, que si dicho empresario revelara datos esenciales del algoritmo (por ej. su código fuente) podría enfrentarse a una demanda de respon-

perjuicio económico para su titular y siempre que se hubiera mantenido en secreto. También aclara que la protección de los secretos empresariales es independiente de la protección por derechos de autor, ya que el derecho de autor protege la expresión concreta del algoritmo (el código fuente o objeto), mientras que el secreto empresarial protege el contenido y el conocimiento técnico o comercial, siempre que no se haga público y se mantenga como confidencial.

⁴⁰⁴ En estos casos, y salvo pacto en contrario, los derechos de explotación del programa corresponden al empresario en virtud del art. 97.4 LPI: «Cuando un trabajador asalariado cree un programa de ordenador, en el ejercicio de las funciones que le han sido confiadas o siguiendo las instrucciones de su empresario, la titularidad de los derechos de explotación correspondientes al programa de ordenador así creado, tanto el programa fuente como el programa objeto, corresponderán, exclusivamente, al empresario, salvo pacto en contrario».

⁴⁰⁵ Como en su momento se explicará, el «responsable del despliegue» es definido en el art.3.4 RIA como la «persona física o jurídica o autoridad, órgano u organismo de otra índole públicos *que utilice un sistema de IA bajo su propia autoridad*, salvo cuando su uso se enmarque en una actividad personal de carácter no profesional».

sabilidad civil del proveedor tecnológico o desarrollador del sistema (titular de los derechos sobre el mismo) reclamándole daños y perjuicios por desvelar dicha información. Por otro lado, que únicamente podría proporcionar información algorítmica (incluso aquella a la que por ley está obligado cuando se trata de decisiones plenamente automatizadas *ex arts. 13.2 f), 14.2 g) y 15.1 h)* en relación con el art. 22 RGPD⁴⁰⁶) cuando contara con dicha información, es decir, cuando el desarrollador de la tecnología le hubiera dado esa información⁴⁰⁷. De ahí la importancia de la articulación y cumplimiento de las obligaciones que impone el art. 13 RIA sobre la información adecuada y suficiente que ha de proporcionarse a los responsables del despliegue que utilicen sistemas de IA de alto riesgo. De hecho, en el Considerando 72 se apunta que el fin de la transparencia y comunicación de información es *ayudar a los responsables del despliegue a cumplir sus obligaciones*.

Por último, podría ocurrir que el responsable del despliegue hubiera obtenido una copia con licencia del algoritmo. En este caso, la doctrina considera que dicho sujeto estaría autorizado en virtud de la Directiva 2009/24/CE a «*observar, estudiar o verificar el funcionamiento de un programa de ordenador con el fin de determinar las ideas y los principios implícitos en cualquier elemento del programa*», ya que estas no están protegidas por los derechos de autor que cubre dicha norma⁴⁰⁸. En consecuencia, aunque no pudiera dar acceso al código fuente (por no ser el titular de sus derechos), sí estaría legitimado para permitir el mero uso del programa a los solos efectos de verificar si el mismo discrimina o no⁴⁰⁹.

⁴⁰⁶ En tales casos, en los que se ha consagrado un *derecho de información reforzado*, el responsable del tratamiento deberá proporcionar al interesado información significativa sobre la lógica aplicada por el algoritmo, en términos comprensibles y al mismo tiempo suficientemente exhaustiva, pero sin necesidad de incluir información sobre los algoritmos utilizados o la revelación de todo el algoritmo.

⁴⁰⁷ Sobre esta cuestión volveremos más adelante al tratar el caso resuelto por la STJUE de 7 de diciembre de 2023, asunto C 634/21 (OQ y Land Hessen, con intervención de SCHUFA Holding AG).

⁴⁰⁸ NAVAS NAVARRO, S., *Daños ocasionados por sistemas de inteligencia artificial...*, cit., p. 47, en base al art. 5.3 de la Directiva 2009/24/CE de 23 de abril sobre protección jurídica de los programas de ordenador, que establece que no constituye infracción de los derechos de autor, la observación, estudio o verificación de su funcionamiento con el fin de determinar las ideas y principios implícitos, no siendo esta excepción derogable por contrato.

⁴⁰⁹ En este sentido, RIVAS VALLEJO, P. «Análisis...», cit., pp. 443-444 basándose en la STJUE Gran Sala, de 2 de mayo de 2012, asunto *SAS Institute Inc contra World Programming Ltd*, razonamiento 50, según el cual, «el licenciataria está autorizado para observar, estudiar o verificar el funcionamiento de un programa de ordenador con el fin de determinar las ideas y los principios implícitos en cualquier elemento del programa». Según establece el Tribunal, en el apartado 59 de la sentencia, «el titular de los derechos de autor sobre un programa de ordenador no puede invocar el contrato de licencia para impedir que quien haya obtenido esa licencia determine las ideas y los principios implícitos en todos los elementos de ese programa cuando realiza las operaciones autorizadas por dicha

Por otro lado, se han alegado también razones de seguridad pública y defensa nacional para denegar el acceso al código fuente de los algoritmos. Así ocurrió, por ejemplo, en el polémico caso del algoritmo *Bosco* utilizado por la Administración española para la concesión del bono social eléctrico a personas vulnerables. Puesto que este caso se analizará con detalle en el apartado 2.5.6.1, baste decir por el momento que los tribunales que conocieron el asunto se basaron en diversos informes técnicos para mantener que revelar el código fuente podría comprometer la seguridad de la información y aumentar la vulnerabilidad ante ataques informáticos, lo que podría afectar adversamente la integridad de datos personales y la seguridad pública⁴¹⁰.

En relación con el acceso al código fuente, el art. 74.13 del RIA posibilita que las *autoridades de vigilancia del mercado* (no el público en general) puedan tener acceso, previa solicitud motivada, si se cumplen dos condiciones: a) que dicho acceso sea necesario para llevar a cabo la evaluación de la conformidad del sistema con los requisitos establecidos en el capítulo III, sección 2; y b) que se hubieran agotado todos los procedimientos de prueba o auditoría y todas las comprobaciones basadas en los datos y la documentación facilitados por el proveedor, o estos fueran insuficientes. Ahora bien, como establece el apartado 14, toda la información o documentación a la que tengan acceso tales autoridades de vigilancia deberá ser tratada de conformidad con las obligaciones de confidencialidad establecidas en el artículo 78. Ello con el fin de proteger los derechos de propiedad intelectual e industrial y la información empresarial confidencial o los secretos comerciales de una persona física o jurídica, incluido el código fuente.

2.4.1.2. Sobre los datos de entrenamiento

No sólo hay que plantearse cuestiones de propiedad intelectual o secreto empresarial sobre el código fuente del algoritmo, sino que otro escollo a la

licencia (...), siempre y cuando no infrinja los derechos exclusivos de ese titular sobre tal programa». Opina la autora que «no precisándose el contexto en el que pueda tener lugar esa observación, bien puede admitirse que la empresa no puede negar el acceso al código fuente a los solos efectos de verificar que no incurre en discriminación cuando se plantee una reclamación de este tenor. Pero no es menos cierto, que igualmente el razonamiento 61 de la sentencia identifica esta situación con el mero uso del programa, sin acceso al código fuente [...]. Es decir, en realidad el TJUE traza una clara frontera entre *estudiar, observar, verificar... y acceder al código fuentes*» (la cursiva es nuestra).

⁴¹⁰ Tanto el Consejo de Transparencia y Buen Gobierno estatal (CTBG) como la Comisión de Garantía del Derecho de Acceso a la Información Pública de Cataluña (GAIP), se han pronunciado en diversas ocasiones sobre el acceso al código fuente llegando la mayoría de las veces (aunque no todas) a soluciones favorables al interesado que después han sido anuladas por los órganos jurisdiccionales. Así ocurrió, por ejemplo, en la resolución RT/253/2021 que fue recurrida por la Comunidad de Madrid y el recurso contencioso-administrativo fue estimado por la Sentencia n° 158/2022 del Juzgado Central Contencioso-Administrativo n° 11.

transparencia, que a su vez dificulta la averiguación del origen de la discriminación algorítmica, es la propiedad de los datos utilizados para entrenar al algoritmo⁴¹¹. Esta cuestión no es baladí por cuanto los expertos en *machine learning* aseveran que lo realmente útil para detectar el sesgo del sistema es el acceso a los conjuntos de datos que han servido de soporte al aprendizaje del algoritmo. Por tanto, en ocasiones no sería suficiente con tener acceso o información del código fuente sino se comparten, además, los datos que lo alimentan.

Así ocurriría en los casos de algoritmos predictivos cuyo aprendizaje se basa precisamente en los datos aportados, a diferencia de lo que ocurriría en los algoritmos donde los datos de alimentación son secundarios. Ejemplo de estos últimos podría ser el de sistemas de IA para solicitar ayudas sociales donde el algoritmo lo único que debe determinar es si los solicitantes cumplen o no los requisitos legales para obtenerla, criterios que a su vez son previamente introducidos por los programadores cuando diseñan el algoritmo. Al tratarse de una decisión reglada cuyo contenido depende de la ley, el algoritmo no influye en el contenido de la decisión ni en su legalidad, sino que es un mero apoyo para su elaboración con el que se trata de facilitar el trabajo humano y evitar errores. Si existen conceptos legales indeterminados no los aplica el algoritmo sino el humano que introduce los datos. No obstante, si estuviera mal diseñado podría producir otro tipo de errores. En estos casos, el control de la legalidad de la decisión habría que hacerlo revisando lo que ha hecho el programador para detectar fallos en alguno de sus pasos. Por tanto, acceder al algoritmo (código fuente o secuencia de programación) arrojaría luz sobre posibles errores a la hora de introducir los criterios exigidos por la norma que llevaron al sistema a adoptar decisiones contrarias a Derecho (por ejemplo, porque se introdujeron requisitos distintos a los previstos en ella)⁴¹².

Por el contrario, en los sistemas de *machine learning* basados en datos, el acceso al código fuente no proporciona al reclamante la información necesaria para demostrar la discriminación algorítmica, sino que normalmente habrá que acceder a las bases de datos para poder detectar el origen del sesgo. Siendo esto así, algunos autores se preguntan si cabría conceder un derecho de acceso a dichos datos a la persona que impugna la decisión del sistema⁴¹³.

Para responder a esta cuestión habría que tener en cuenta si dichos datos de entrenamiento son públicos o privados, personales o no personales, ya que, mientras el uso de datos no personales no está protegido por el RGPD,

⁴¹¹ RIVAS VALLEJO, P. «Análisis...», *cit.*, pp. 448 y ss. Sobre el derecho de acceso a los datos de entrenamiento de sistema, puede verse ampliamente, LÓPEZ-TARRUELLA MARTÍNEZ, A., *Propiedad intelectual e innovación basada en los datos*, Navarra, 2021, pp. 192 y ss.

⁴¹² Así lo explica HUERGO LORA, A., «Una aproximación a los algoritmos ...», *cit.*, p. 72.

⁴¹³ RIVAS VALLEJO, P., «Análisis ...», *cit.*, p. 448

el uso no autorizado de datos personales puede conllevar responsabilidades y sanciones e incluso, en algunos casos, la propia nulidad de la herramienta y por ende de la decisión algorítmica⁴¹⁴.

El RGPD contempla el derecho de la persona afectada por la decisión a acceder a *sus propios datos personales* y a obtener información sobre su tratamiento⁴¹⁵, pero no a otros datos que pudieran afectarle, como serían los datos de terceras personas utilizados en el análisis (por ejemplo, los datos históricos de la empresa en base a los cuales el algoritmo aprende y toma decisiones). Además, cuando el interesado es objeto de una decisión plenamente automatizada con efectos jurídicos o similares sobre él y basada en sus datos personales, también se recoge en el art. 14.1 RGPD el derecho a ser informado sobre la lógica aplicada y la importancia o consecuencias que puede tener para él dicho tratamiento. Precisamente en base a este derecho un sector doctrinal considera que se podría defender el acceso del interesado a los datos empleados por el sistema para la toma de decisión⁴¹⁶. Pero en cualquier caso, esos datos deberían ser objeto de anonimización, pseudoanonimización, cifrado, desagregación, aleatorización o cualquier otra medida de seguridad, a fin de preservar el derecho a la privacidad (arts. 32 y ss RGPD).

Aun manteniendo la existencia de tal derecho en el RGPD, el mismo no implicaría la obligación del titular de los datos de dar acceso al interesado a todos los datos de entrenamiento del sistema, sino sólo a aquellos que se utilizaron para adoptar la decisión respecto de la que se solicita su revisión. Es más, en virtud del art. 63 RGPD el responsable del tratamiento podría negarse a otorgar un acceso general a los datos y solicitar que el interesado especifique la información o actividades de tratamiento a que se refiere la solicitud.

En cualquier caso, sin perjuicio de lo anterior, debe partirse de que los datos que se emplean para entrenar estos sistemas no suelen ser datos personales, sino anónimos o anonimizados. Por ende, aunque puedan contener características o rasgos personales, si no identifican a personas concretas no estarían protegidos por el RGPD⁴¹⁷.

⁴¹⁴ Sobre esto nos remitimos al apartado 2.5.2 del Capítulo III y el apartado 3.2 del Capítulo V.

⁴¹⁵ Art.15 RGPD y art. 13 de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (en adelante, LOPDPGDD).

⁴¹⁶ En este sentido, manifiesta LÓPEZ-TARRUELLA MARTÍNEZ, A., *op.cit.*, p. 196, que el acceso a los datos podría venir de la mano del llamado derecho de explicabilidad o del derecho a una explicación. No obstante, como sigue explicando, este derecho presenta dos problemas: por un lado, su incierto refrendo legal; y por otro, su incompatibilidad con la protección de los datos por derechos de propiedad intelectual y secreto empresarial.

⁴¹⁷ Como afirma la doctrina, los datos personales juegan un papel secundario en las decisiones automatizadas adoptadas por estos sistemas inteligentes. Por todos, HUERGO LORA, A., «Una aproximación a los algoritmos ...», *cit.*, p. 53.

Siendo esto así, para determinar el posible acceso del interesado a los mismos habría que analizar la titularidad sobre dichas bases de datos. En este sentido, aunque según los arts. 133 y siguientes de la LPI, las bases de datos también pueden ser objeto de derechos de propiedad intelectual, hay quien mantiene que la excepción recogida en el art 135.c)⁴¹⁸ permitiría que en el ámbito de un procedimiento judicial se pudiera acceder a las bases de datos que alimentan el algoritmo para proteger el interés legítimo del afectado por una decisión automatizada, *a los solos efectos de conocer si existe un sesgo en la fuente de tal decisión*⁴¹⁹. Por tanto, siempre y cuando el acceso estuviese limitado a este fin y la base no contuviese datos identificativos de personas concretas, el interesado podría llegar a acceder a esa información.

Al margen de los supuestos que puedan quedar cubiertos por dicha excepción, lo cierto es que hasta fechas recientes ninguna norma del acervo comunitario obligaba expresamente a ofrecer acceso a los datos de entrenamiento para investigar posibles errores del sistemas o sesgos discriminatorios⁴²⁰. Para LÓPEZ-TARRUELLA, cuya opinión compartimos, resultaba paradójico que no existiera en el Derecho de la UE una norma de propiedad intelectual que facilitara la investigación de sesgos algorítmicos en general, o de aquellos que pudieran afectar a derechos fundamentales en particular, sobre todo teniendo en cuenta que para las instituciones europeas la transparencia y la explica-

⁴¹⁸ El art. 135 LPI recoge una serie de excepciones a los derechos *sui generis* sobre las mismas y dispone que «el usuario legítimo de una base de datos, sea cual fuere la forma en que ésta haya sido puesta a disposición del público, podrá, sin autorización del fabricante de la base, extraer y/o reutilizar una parte sustancial del contenido de la misma, en los siguientes casos: a) Cuando se trate de una extracción para fines privados del contenido de una base de datos no electrónica. b) Cuando se trate de una extracción con fines ilustrativos de enseñanza o de investigación científica en la medida justificada por el objetivo no comercial que se persiga y siempre que se indique la fuente. c) *Cuando se trate de una extracción y/o reutilización para fines de seguridad pública o a efectos de un procedimiento administrativo o judicial*».

⁴¹⁹ Así, RIVAS VALLEJO, P., «Análisis ...», *cit.*, p. 449.

⁴²⁰ Opina LÓPEZ-TARRUELLA MARTÍNEZ, A., *op.cit.*, p. 199, que esta laguna podía haberse cubierto mediante una clarificación jurisprudencial o legislativa del derecho de explicabilidad y de su alcance. También considera que podría invocarse el art. 1.2 c) de la Directiva 2016/943 para impedir que el titular de los datos hiciera valer la protección por secreto empresarial para negarse a conceder acceso, al menos en aquellos supuestos en los que el solicitante fuera una autoridad pública.

En el acervo comunitario, la fallida *Propuesta de Reglamento sobre un régimen de responsabilidad civil en materia de inteligencia artificial, de 20 de octubre de 2020*, sí que contemplaba un derecho de acceso a los datos generados por el sistema de alto riesgo (fuesen datos personales o de otro tipo, especialmente técnicos) tanto a la víctima como al operador responsable. Sobre esto el art. 10.2 disponía lo siguiente: «Un operador considerado responsable podrá utilizar los datos generados por el sistema de IA para demostrar la negligencia concurrente de la persona afectada, de conformidad con el Reglamento (UE) 2016/679 y otras leyes en materia de protección de datos relevantes. La persona afectada también podrá usar esos datos con fines probatorios o aclaratorios en la demanda por responsabilidad civil».

bilidad eran componentes esenciales de la IA centrada en el ser humano⁴²¹. Para dicho autor, en estos supuestos estaría justificado el acceso a los datos, e incluso devendría una cuestión de orden público, pero habría que conceder ese acceso bajo ciertas salvaguardas que garantizaran los legítimos intereses del titular de los datos y la protección de información confidencial⁴²².

Pues bien, en fechas recientes se han aprobado diversas normas que garantizan un derecho de acceso a los datos e información algorítmica. A ellas haremos referencia en el apartado sobre las medidas para hacer frente a la opacidad de la IA. Baste citar por el momento, a modo de ejemplo, que la nueva *Ley de Datos* que entró en vigor el 11 de enero de 2024 y será aplicable en septiembre de 2025⁴²³, establece en su art. 4 el derecho del usuario de un producto o servicio digital contratado a acceder a los datos que el mismo genere. En opinión de NAVAS NAVARRO, dicha norma ampara el acceso a los datos a los efectos de una posible futura litigación por daños⁴²⁴.

Asimismo, las últimas Propuestas de Directivas de la Comisión Europea en materia de responsabilidad por daños ocasionados por sistema de IA⁴²⁵, también introducen importantes medidas a fin de remover los obstáculos para acceder a la información algorítmica mediante el reconocimiento de un *derecho*

⁴²¹ Vid., por ejemplo, las *Directrices éticas sobre IA*.

⁴²² LÓPEZ-TARRUELLA MARTÍNEZ, A., *op.cit.*, pp. 195-196. Considera el autor que regulando el acceso a los datos bajo ciertas condiciones se evitaría desincentivar a las organizaciones a invertir en el desarrollo de herramientas de IA ante la posibilidad de verse obligados a compartir los conjuntos de datos en los que se basa su ventaja competitiva. Vid., p. 201. Precisamente este fundamento es el esgrimido en el RIA para justificar el tratamiento excepcional de categorías especiales de datos a través de sistemas de IA. Así, evitar la discriminación que podría provocar el sesgo de los sistemas de IA se considera una cuestión de orden público esencial que permitiría que los proveedores trataran ese tipo de datos. De ahí que su art. 10.5 disponga que «*En la medida en que sea estrictamente necesario para garantizar la vigilancia, la detección y la corrección de los sesgos asociados a los sistemas de IA de alto riesgo, los proveedores de dichos sistemas podrán tratar las categorías especiales de datos personales [...] ofreciendo siempre las salvaguardias adecuadas para los derechos y las libertades fundamentales de las personas físicas, lo que incluye establecer limitaciones técnicas a la reutilización y la utilización de las medidas de seguridad y protección de la privacidad más recientes, tales como la seudonimización o el cifrado, cuando la anonimización pueda afectar significativamente al objetivo perseguido*».

⁴²³ Reglamento (UE) 2023/2854 del Parlamento Europeo y del Consejo, de 13 de diciembre de 2023, sobre normas armonizadas para un acceso justo a los datos y su utilización, y por el que se modifican el Reglamento (UE) 2017/2394 y la Directiva (UE) 2020/1828 (Reglamento de Datos).

⁴²⁴ NAVAS NAVARRO, S., *Daños ocasionados por sistemas de inteligencia artificial. Especial atención a su futura regulación*, Granada, 2022, p. 71.

⁴²⁵ Una de esas propuestas, en concreto, la relativa a la modernización del régimen de responsabilidad por productos defectuosos, ha sido recientemente aprobada y publicada en el DOUE núm. 2853, de 18 de noviembre de 2024: *Directiva (UE) 2024/2853 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, sobre responsabilidad por los daños causados por productos defectuosos y por la que se deroga la Directiva 85/374/CEE del Consejo*.

a la exhibición de pruebas a favor de las víctimas⁴²⁶. Al análisis de este expediente dedicaremos un apartado concreto de esta obra.

2.4.2. La opacidad técnica o consustancial al algoritmo

Además de los factores jurídicos que pueden justificar la opacidad de estos sistemas, la misma puede deberse también a la propia complejidad intrínseca del algoritmo. Dicha opacidad aumenta cuando se emplean algoritmos de autoaprendizaje no supervisado y profundo que gozan de gran autonomía decisoria y cuya capacidad predictiva (altamente precisa por otro lado) reside en establecer correlaciones tan sumamente complejas que sus resultados pueden ser ininteligibles para el ser humano a partir de los datos que se aportan al modelo. Este efecto, conocido como «caja negra» o *black box*, hace referencia a la imposibilidad de identificar las razones que han llevado al sistema de IA a adoptar sus decisiones. Por ello, el uso de esta tecnología plantea un problema estructural en la medida en que su esencia misma choca con las normas que exigen transparencia, trazabilidad y posibilidad de identificar los procesos de toma de decisiones de los algoritmos.

Respecto a este tipo de sistemas donde la lógica computacional es elusiva y opaca, ni siquiera la apertura de la caja negra y el acceso al código fuente del algoritmo garantizaría la comprensibilidad o inteligibilidad de los resultados⁴²⁷. Por un lado, porque conocer el código fuente únicamente permitiría saber la utilidad pretendida por el algoritmo presente en su diseño (por ejemplo, encontrar al mejor candidato para un puesto de trabajo o determinar cuánto dinero costará un paciente, así como la secuenciación de pasos o instrucciones para conseguir ese objetivo), pero como señala RIVAS VALLEJO, no permitiría conocer «el proceso de gestión de capas de datos que debe salvarse para llegar a sus “entrañas”, inaccesible por la propia dinámica del aprendizaje profundo»⁴²⁸. Por otro lado, porque con carácter general para comprender el algoritmo y los datos utilizados por la IA hacen falta una capacidad analítica y unos conocimientos técnicos que pueden ser excesivamente costosos

⁴²⁶ Ese derecho se ha acabado consagrando en el art. 9 de la Directiva 2024/2853.

⁴²⁷ Así lo afirman, entre otros muchos, SOLAR CAYÓN, J.I., *op.cit.*, p. 149; EVANGELIO LLORCA, R., «Causalidad y responsabilidad civil por daños ocasionados por sistemas de inteligencia artificial: las presunciones de causalidad en las propuestas normativas de la UE», en *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, coord. por N. Álvarez Lata; J.M. Busto Lago (pr.), 2024, p. 561. No obstante, como mantiene NAVAS NAVARRO, S., «Sistemas expertos basados en inteligencia artificial y responsabilidad civil. Algunas cuestiones controvertidas», *Diario La Ley*, 13 de Diciembre de 2019, LA LEY 14980/2019, p. 7/16, si bien el acceso a la *black box* no será la panacea para la víctima, quizá sí que sea de utilidad para fijar indicios.

⁴²⁸ RIVAS VALLEJO, P. «Análisis...», *cit.*, p. 438.

para las víctimas. E incluso en los sistemas más complejos podría ocurrir que ni siquiera los técnicos en ciencias de la computación pudieran ofrecer una explicación acerca de cómo o por qué se ha llegado a un concreto resultado, ya que, en tales casos, «los diseñadores dejan de ser partícipes en los procesos decisionales que adopta su creación, la cual cobra vida propia siendo opaca al exterior»⁴²⁹.

Este problema estructural hace que la propia idea de la transparencia resulte cuestionable para algunos autores e incluso se hable de la «falacia de la transparencia»⁴³⁰, ya que no siempre que se consigue acceder a la caja negra quedan salvaguardados adecuadamente los derechos de los interesados. Especialmente a los efectos de la asignación de responsabilidad, el acceso a la información de las cajas negras puede no ser suficiente para lograr una conexión entre el autor, el algoritmo y sus consecuencias, que sea lo suficientemente sustancial como para que alguien pueda considerarse responsable de sus efectos perjudiciales⁴³¹.

Las instituciones europeas han venido advirtiendo de este *handicap* en diferentes documentos, destacando que dicho efecto de caja negra podría llegar a impedir que los afectados obtuvieran una indemnización por los daños causados por aplicaciones de IA autónomas⁴³². Piénsese que dicha opacidad dificultaría en gran medida (e incluso imposibilitaría) la averiguación de la causa u origen de un determinado resultado discriminatorio (por ejemplo, si el fallo está en una correlación a la que ha llegado el sistema tras su aprendizaje autónomo, si está en los datos aportados, si reside en su propio diseño, etc.) y por ende la determinación de los posibles responsables de los daños ocasionados. Pero es que, además, incluso cuando se pudiera detectar la causa del problema, la misma podría variar o desaparecer con el tiempo teniendo en cuenta que este tipo de algoritmo está en constante evolución⁴³³.

⁴²⁹ SUBERBIOLA GARBIZU, I., *op.cit.*

⁴³⁰ SOLAR CAYÓN, J.L., *op.cit.*, p. 148.

⁴³¹ GONZÁLEZ VALVERDE, A., *op.cit.*, pp. 1263-1264.

⁴³² Vid., por ejemplo, el *Informe sobre las repercusiones en materia de seguridad y responsabilidad civil de la IA, el internet de las cosas y la robótica*, de 19 de febrero de 2020, de la Comisión.

⁴³³ Precisamente por ello, como en su momento se explicará, se venía abogando por la necesidad de que esta tecnología contara con un sistema que permitiera registrar información sobre su funcionamiento, esto es, datos relativos a su movimiento, actuación y decisión («logging by design» o registro desde el diseño) a modo de *caja negra* como la que se encuentra hoy en día en aviones, trenes o drones. Entre otros lo defendían, NAVAS NAVARRO, S., «La perspectiva de género ...», *cit.*, pp. 8 y 11/23, basándose en el Informe del *Expert Group on Liability and New Technologies, Liability*; o SORIANO ARNANZ, A., «La aplicación del marco jurídico europeo...», *cit.*, p. 81, para quien en los sistemas que se van actualizando, es necesario que los operadores guarden una copia de cada versión del programa utilizado en los procesos de toma de decisiones.

Por estos motivos, hay quien opina que debería demorarse el empleo de este tipo de sistemas de aprendizaje profundo en los que no se puede garantizar el respeto de los derechos fundamentales, aun a costa de sacrificar sus ventajas (principalmente, su mayor nivel de precisión)⁴³⁴. De hecho, la ininteligibilidad de algunos algoritmos de caja negra ha frenado su empleo en sectores como la medicina, limitando su uso como soporte o apoyo en la toma de decisiones clínicas⁴³⁵. Es más, como explica RUBÍ PUIG⁴³⁶, se ha defendido que a la hora de valorar la culpa para determinar la responsabilidad por el uso de sistemas de *machine learning* en la prestación de servicios médicos, debería examinarse la explicabilidad de los modelos, de forma que podría resultar negligente para un médico utilizar modelos opacos, salvo cuando quedara acreditado que la tasa de corrección de sus predicciones fuera significativamente superior a la ofrecida por modelos transparentes. Para dicho autor, la aplicación de este criterio podría incentivar a los desarrolladores de sistemas de IA a diseñar modelos más transparentes habida cuenta de que su opacidad podría conllevar que su empleo por los usuarios (por ejemplo, un hospital que lo implementa) pudiera considerarse negligente por un tribunal. Por su parte los usuarios, conociendo que podrían devenir responsables por el uso de sistemas de caja negra o indescifrables, tratarían de evitar su utilización en pro de otros modelos transparentes y de similar precisión, en caso de haberlos⁴³⁷.

⁴³⁴ En el ámbito europeo, el Informe *AI in the UK: ready, willing and able?*, elaborado por el Comité sobre inteligencia artificial de la Cámara de los Lores británica, recomienda que el empleo de sistemas de IA basados en redes neuronales profundas que puedan tener un impacto sustancial en la vida de los individuos se demore hasta que se encuentren soluciones técnicas que permitan su comprensibilidad plena (ap. 105). Accesible en: <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm>. En relación con esto, considera ALCOLEA AZCÁRRAGA, C., *op.cit.*, p. 8/31, la posibilidad de extrapolar al ámbito tecnológico el *principio de precaución* aplicable en Derecho ambiental. Con el fin de controlar las consecuencias imprevisibles de los sistemas de IA y evitar daños irreparables, esgrime que podría adoptarse el principio de precaución algorítmica, por cuanto el ámbito tecnológico tiene también un importante potencial dañino para la humanidad en su conjunto. En base a dicho principio de precaución concluye el autor que, aunque algunos «autores consideran que la aplicación de los algoritmos a las facultades discrecionales de la Administración pública aportaría un baremo nuevo que alejaría la arbitrariedad, podría parecer que su uso sin una completa comprensión de cómo funcionan o por qué toman las decisiones que toman debería postergarse a un momento en que efectivamente se disponga de ese control y posibilidad de fiscalización». En definitiva, y en la línea apuntada por SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, pp. 94-5, creemos que en esta materia es crucial decidir si se debe dar prioridad a la eficiencia y mayor precisión de la tecnología *black box* sobre los criterios necesarios para garantizar que una decisión está fundada en Derecho, o si, por el contrario, derechos fundamentales como la tutela judicial efectiva y todo lo que implica, no pueden ceder ante la IA.

⁴³⁵ LAZCOZ MORATINOS, G., «Análisis jurídico de la toma de decisiones algorítmica...», *cit.*, p. 293.

⁴³⁶ «Retos de la inteligencia artificial ...», *cit.*, p. 65.

⁴³⁷ Por tanto, si existiendo modelos transparentes de igual fiabilidad, se utilizaran modelos opacos y con ellos se causasen daños, se debería responder por su uso. Ahora bien, a falta de modelos

Frente a la cautelosa postura que defiende la demora en la aplicación de modelos de *black box*, la opinión mayoritaria aboga por la utilización de tales sistemas adoptando, eso sí, un enfoque de «explicabilidad» sobre la mera transparencia técnica. De hecho, existe actualmente una tendencia a desarrollar modelos algorítmicos más sencillos, transparentes y explicables que respondan mejor a las exigencias jurídicas y generen mayor confianza en los usuarios, que ha dado lugar a un campo de estudio e investigación denominado *Explainable Artificial Intelligence* o *Interpretable Machine Learning*, a través del que se pretende encontrar un equilibrio adecuado entre la optimización de la capacidad predictiva de los algoritmos y su explicabilidad. Se trata de integrar la *transparencia desde el diseño* de forma que los sistemas basados en IA se diseñen de tal manera que los usuarios puedan comprender el porqué de sus acciones.

El *grado* de explicabilidad dependerá en gran medida del contexto en el que se enmarquen estos sistemas y la gravedad de las consecuencias derivadas de un resultado erróneo o impreciso (por ejemplo, un sistema de IA que genere unas recomendaciones de compra poco acertadas no despertará excesivas preocupaciones desde el punto de vista ético, a diferencia de uno dirigido a evaluar la concesión de la libertad condicional al condenado), y debe ser *suficiente* para poder impugnar los resultados y solicitar un resarcimiento por los daños que se generen⁴³⁸. Además, en la medida de lo posible, la explicabilidad del proceso de toma de decisiones algorítmico debe estar adaptada a las personas afectadas⁴³⁹.

Para COTINO HUESO, las variables que podrían condicionar el grado o intensidad de explicabilidad y transparencia exigible serían las siguientes: en primer lugar, el impacto que tiene el algoritmo en la decisión, el resultado y el ciudadano; en segundo lugar, el grado de autonomía del sistema en la toma de decisiones (es decir, hasta qué punto se garantiza la participación humana); y por último, el tipo y la complejidad del algoritmo. Así, cuanto mayor sea el impacto de las decisiones algorítmicas en los ciudadanos y mayor sea

transparentes igualmente precisos –o con tasa de error mucho mayor– podría plantearse si se debería asumir responsabilidad si el sistema causara daños.

⁴³⁸ Así lo afirma el Parlamento Europeo en la Propuesta de Reglamento que acompañaba a su Resolución sobre aspectos éticos de la IA, de 20 de octubre de 2020. Vid., Considerando 19, donde finaliza diciendo que «La auditabilidad, la trazabilidad y la transparencia deben tratar cualquier posible ininteligibilidad de estas tecnologías».

⁴³⁹ Vid., Comunicación de la Comisión Europea «Generar confianza en la inteligencia artificial centrada en el ser humano» (COM/2019/168 final), p. 6 o *Directrices éticas* del Grupo de Expertos, p. 22, donde se apunta que la explicación «debería ser oportuna y adaptarse al nivel de especialización de la parte interesada (que puede ser una persona no experta en la materia, un regulador o un investigador)».

la autonomía, opacidad y complejidad del sistema de IA, mayores niveles de transparencia y explicabilidad se tendrán que exigir⁴⁴⁰.

Respecto a los algoritmos de *black box*, en los que no es posible obtener una explicación acerca de cómo se ha generado un resultado determinado o qué combinación de *inputs* han contribuido al mismo, la explicabilidad debería traducirse en medidas como la *trazabilidad*, la *auditabilidad* o la *comunicación transparente de las capacidades del sistema* para que las personas afectadas puedan comprender cuál es la base de sus acciones o decisiones⁴⁴¹. Es decir, en tales casos es necesario adoptar otro tipo de estrategias de control con el fin de desarrollar sistemas algorítmicos explicables desde un punto de vista práctico. Más allá de la apertura de la caja negra (que como hemos visto no siempre será la panacea para la víctima), dichas estrategias deben ir dirigidas al escrutinio de los datos de entrada y salida y al establecimiento de una serie de garantías en el proceso de diseño, entrenamiento e implementación de los sistemas y de procedimientos de monitorización y supervisión de los resultados. Garantías y procedimientos que además deberán someterse a auditorías. Como se expondrá en su momento, muchas de estas estrategias han sido recogidas en el nuevo RIA.

2.5. MEDIDAS PARA HACER FRENTE A LOS PROBLEMAS DE OPACIDAD ALGORÍTMICA EN LA TOMA DE DECISIONES A TRAVÉS DE SISTEMAS DE IA

2.5.1. Estado de la cuestión

Hasta hace poco tiempo el único instrumento para hacer frente a la opacidad de los sistemas automatizados de decisiones, basados o no en IA, era el

⁴⁴⁰ Así lo entiende COTINO HUESO, L., «Transparencia de la inteligencia artificial pública...», cit. También en «Qué concreta transparencia e información de algoritmos e inteligencia artificial es la debida» *Revista Española de la Transparencia*, Núm. 16, 2023, pp.17-63, accesible en file:///C:/Users/ASUS/Downloads/Dialnet-QueConcretaTransparenciaEInformacionDeAlgoritmosEI-8913030.pdf

⁴⁴¹ La Comisión Europea ha señalado en diversos documentos que en relación a aquellas aplicaciones de IA que entrañan un riesgo elevado, el marco regulador podrá exigir la conservación de registros exactos sobre el conjunto de datos utilizados para entrenar el algoritmo, una descripción de sus principales características y del modo en que se escogió, así como de documentación sobre las metodologías de programación y entrenamiento y los procesos y las técnicas utilizados para construir, probar y validar los sistemas; además de obligar a facilitar información clara de sus capacidades y limitaciones, objetivo y condiciones en las que se espera que funcione según lo previsto. Incluso considera el diseño de *sistemas mixtos* que combinen redes neuronales profundas y razonamiento simbólico (esto es, reglas creadas mediante intervención humana) como una solución técnica que ayude a mejorar la capacidad de explicar los resultados de la IA. Vid., *Directrices éticas* p. 16 y también el *Libro Blanco sobre la inteligencia artificial* (COM (2020) 65 final), pp. 23-25.

RGPD. Sin embargo, en los últimos años se han presentado otras propuestas regulatorias por las instituciones europeas y se han ido aprobando paulatinamente normas a nivel interno y comunitario, tendentes a garantizar en mayor o menor medida la transparencia algorítmica y el derecho a una explicación, esto es, el derecho a conocer cómo funciona el concreto proceso de toma de decisión con indicación de los datos específicos de la persona en base a los cuales se ha tomado una decisión o se ha hecho un perfilado. El objetivo de todas ellas es eliminar, o al menos mitigar, la opacidad tanto en el *uso* como en el *contenido* de la información sobre el algoritmo (es decir, cómo funciona, cuáles son las instrucciones o reglas que conducen a la toma de decisión, cuáles son las variables introducidas y el peso otorgado a cada una de ellas, etc), y se asientan sobre una serie de principios como el de transparencia desde el diseño, minimización de sesgos y rendición de cuentas.

Además, contamos con pronunciamientos judiciales y de órganos administrativos que se posicionan a favor de garantizar un nivel de transparencia algorítmica mucho mayor y obligaciones de información más exigentes en los casos en que los sistemas de IA para la automatización de decisiones se utilizan en el ámbito público⁴⁴².

Por último, la relativa inexistencia de normas que obligaran a las empresas a dar acceso al código fuente de sus algoritmos o a divulgar sus características frente a la reclamación de un particular, ha llevado también a la doctrina a proponer soluciones para garantizar la transparencia algorítmica sin afectar a los derechos de propiedad intelectual y los secretos comerciales. De hecho, encontrar un equilibrio entre la propiedad intelectual de ciertos métodos de procesamiento y la necesidad de garantizar la transparencia, imparcialidad y equidad algorítmica, ha sido considerado por las instituciones europeas como

⁴⁴² Además de la sentencia de 5 de febrero de 2020 del Tribunal de lo Civil de la Haya dictada en el caso del algoritmo Syri, también se obligó a dar acceso al código fuente de diferentes sistemas automatizados empleados por el sector público en los siguientes asuntos: Tribunal Regional Administrativo del Lazio-Roma, Sección III bis, Sentencias N° 3769, de 22 de marzo de 2017 y N° 10964, de 13 de septiembre de 2019; Comisión Francesa de Acceso a los Documentos Administrativos, Decisiones N° 20144578, de 8 de enero de 2015 y N° 20180276, de 19 de abril de 2018. En nuestro país, la Resolución de 21 de septiembre de 2016 de la Comisión de Garantía del Derecho de Acceso a la Información Pública (GAIP), que estima las Reclamaciones 123/2016 i 124/2016 (acumuladas) sobre el algoritmo matemático que determina la selección de los miembros de los tribunales correctores de las pruebas de acceso a la Universidad (PAU) y DNI de los candidatos presentados, y su Resolución 200/2017, de 21 de junio relacionada con las anteriores. Por lo que se refiere al CTBG, las resoluciones R/0058/2021 de 20 de mayo, sobre el acceso al algoritmo para calcular pensiones de la TGSS; RT/253/2021 de 19 de noviembre y RT/ 748/2021 de 10 de enero 2022 sobre acceso al código fuente de una aplicación utilizada para el sorteo de tribunales asociados a procesos selectivos para educación en Madrid, y la Resolución del Expediente 551-2023 sobre el acceso a información del Sistema de Seguimiento Integral de los casos de Violencia de Género (Sistema VioGén).

un objetivo imprescindible desde que comenzaron a trabajar sobre la IA hasta sus más recientes y relevantes propuestas normativas.

Por lo que se refiere al RIA, se establece en su art. 78 que todas las partes implicadas en la aplicación del Reglamento deberán respetar la confidencialidad de la información y los datos que obtengan en el ejercicio de sus funciones, de conformidad con el Derecho nacional o de la Unión, y deberán desempeñar sus funciones y actividades de manera que se protejan, en particular, los derechos de propiedad intelectual e industrial, la información empresarial confidencial y los secretos comerciales de una persona física o jurídica, incluido el código fuente, salvo en los casos mencionados en el artículo 5 de la Directiva (UE) 2016/943 sobre secretos comerciales⁴⁴³.

También las últimas Propuestas de Directivas de la Comisión Europea en materia de responsabilidad por daños ocasionados por sistema de IA abogan por la necesidad de conseguir este equilibrio de intereses. Concretamente, el Considerando 20 de la PDRC dispone que los órganos jurisdiccionales sólo deberán ordenar la exhibición de pruebas cuando sea *necesario y proporcionado* para sustentar la demanda, añadiendo que la información que se suministre se limitará a la *estrictamente necesaria para demostrar el incumplimiento de un requisito fijado por la futura Ley IA*. Todo ello con el fin de salvaguardar los intereses legítimos de todas las partes implicadas. Y respecto a los secretos comerciales, dispone también que los órganos jurisdiccionales nacionales deben estar facultados para adoptar medidas específicas (como por ejemplo, restringir el acceso a los documentos que contengan secretos comerciales y el acceso a las audiencias o los documentos y sus transcripciones a un número limitado de personas) para garantizar la confidencialidad de los mismos durante y después del proceso, al tiempo que se logra un equilibrio justo y proporcionado entre el interés del poseedor del secreto comercial en mantener el secreto y el interés de la persona perjudicada. Al decidir sobre tales medidas, deben tener en cuenta la necesidad de garantizar el derecho a la tutela judicial efectiva y a un juez imparcial, los intereses legítimos de las partes y, en su caso, de terceros, así como el perjuicio que pudiera ocasionarse a cualquiera de las partes o, en su caso, a terceros, como consecuencia de que se acuerden o no dichas medidas.

Ya en su articulado, el art. 3.4, párrafos segundo y tercero, pretenden lograr ese equilibrio entre los derechos del demandante y la necesidad de garantizar que la exhibición de pruebas esté sujeta a garantías que protejan los intereses legítimos de todas las partes interesadas, como los secretos comercia-

⁴⁴³ Como sujetos pasivos de esta obligación se menciona expresamente en dicho precepto a la Comisión, las autoridades de vigilancia del mercado, los organismos notificados y cualquier otra persona física o jurídica que participe en la aplicación del Reglamento. Vid. también el Considerando 167.

les o la información confidencial. Y el párrafo cuarto tiene por objeto garantizar que la persona sujeta a la orden de exhibición o conservación de pruebas tenga remedios procesales a su disposición en respuesta a dichas órdenes.

Todas estas medidas legales, posicionamientos jurisprudenciales y propuestas doctrinales dirigidas a luchar contra algunos de los principales obstáculos que plantea la discriminación algorítmica como son su detección y prueba, serán expuestas en este apartado.

2.5.2. El Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, de protección de datos (RGPD)

La relevancia de esta norma y el hecho de que hasta hace relativamente poco tiempo fuera el único marco jurídico para hacer frente a los riesgos del *machine learning* en la toma de decisiones, nos lleva a analizarla de forma prioritaria al resto de iniciativas que las instituciones europeas han ido presentando (y en su caso, aprobando) desde que comenzaron a regular la IA.

El RGPD establece una serie de garantías y derechos para hacer frente a la opacidad algorítmica cuando el interesado es objeto de una decisión plenamente automatizada basada en sus datos personales y con efectos jurídicos o similares en él: por un lado, los arts.13.2 f) y 14.2 g) obligan a informarle sobre la *existencia de decisiones automatizadas*, incluida la elaboración de perfiles, y al menos en tales casos, a darle *información significativa sobre la lógica aplicada, así como la importancia y las consecuencias que dicho tratamiento puede tener para su persona*⁴⁴⁴. Por tanto, se consagra un derecho de información reforzado o ampliado exigible cuando concurren los presupuestos mencionados; por otro lado, el art. 22.3 establece que en tales casos el responsable del tratamiento adoptará las medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado, y, como mínimo, el *derecho a obtener intervención humana* por parte del responsable, a *expresar su punto de vista* y a *impugnar la decisión* que le afecta.

No hay unanimidad acerca de si el RGPD consagra el que se ha venido a denominar como «derecho de explicabilidad algorítmica» o «derecho a abrir la caja negra de los algoritmos». Mientras el derecho de información *ex ante* está adecuadamente regulado en el RGPD –como se deriva de sus arts. 5.1, 13, 14 o 15– y parece claro que opera con anterioridad a que el sistema adopte la decisión o haga el perfilado, se discute su existencia *ex post*, esto es, después de elaborada la decisión o el perfil de la persona. La controversia se basa, prin-

⁴⁴⁴ Lo mismo se recoge en el contexto del derecho de acceso del art. 15.1.h).

cialmente, en que dicho derecho no se reconoce en el texto articulado sino simplemente en uno de sus considerandos, en concreto en el Considerando 71, según el cual, el afectado por la decisión automatizada goza del «derecho a obtener intervención humana, a expresar su punto de vista, a recibir una explicación de la decisión tomada después de tal evaluación y a impugnar la decisión»⁴⁴⁵.

Partiendo de que los considerandos no son en sí mismos normas, si la concreta decisión o el perfilado adoptados por el sistema resultan discriminatorios y causan daños a la persona, no hay consenso acerca de si el interesado tendría derecho a que se le informara sobre la *lógica decisional* (distinta de la *lógica aplicada* a la que se refieren los art. 13, 14 y 15⁴⁴⁶), es decir, a que se le explicara el proceso decisorio que ha conducido a que el sistema arrojará ese resultado, así como si este derecho incluiría, a su vez, el derecho a acceder a los datos de entrenamiento del sistema. La cuestión no es baladí si se tiene en cuenta que la solicitud de un interesado para investigar posibles errores o ses-

⁴⁴⁵ Sobre esta controversia doctrinal, puede verse LÓPEZ-TARRUELLA MARTÍNEZ, A., *op.cit.*, pp. 197 y ss, y doctrina citada en notas 718 y 719. También sobre el asunto, NUÑEZ SEOANE, J., «El derecho a la información y acceso al funcionamiento de los algoritmos que tratan datos personales», en *La regulación de los algoritmos*, coord. por Gustavo Manuel Díaz González y Alejandro José Huergo Lora (dir.) 2020, pp.307-308.

⁴⁴⁶ Diferencia con claridad entre lo que sería la lógica aplicada y la lógica decisional, GIL MEMBRADO, C., «Daños producidos por la IA: la opacidad del algoritmo y el efecto de caja negra», en N. Álvarez Lata/J.M. Busto Lago, *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, 2024, p. 521. La primera estaría reconocida en los arts. 13 y 14 RGPD y obligaría a informar sobre el sistema de IA en general, mientras que la segunda obligaría a abordar qué proceso decisorio ha conducido a un determinado resultado de salida. En su opinión, el RGPD sólo reconoce la primera (esto es, el derecho a conocer la lógica aplicable *—ex ante—*, pero no el derecho a conocer la explicabilidad *—ex post—*). Para MEDINA GUERRERO, M., «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*, n° 49, 2022, p. 161, el deber de proporcionar «información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas» previsto en los arts. 13.2.f) y 14.2.g) RGPD, sólo puede traducirse en la práctica en una explicación *ex ante* sobre la *funcionalidad del sistema* en la medida en que debe llevarse a cabo en el momento en que se obtengan los datos personales (ex art. 13.2) o en un plazo muy breve (ex art. 14.3.a). Al realizarse la notificación antes de que el sistema decisional haya adoptado la solución concreta, la información que debe dar el responsable no puede relacionarse con el resultado de ninguna decisión específica, sino que deberá centrarse en el sistema considerado como un todo abstracto. Por lo que, como es lógico, en esta fase la obligación de información no puede traducirse en la exigencia de asegurar la explicabilidad o la fundamentación de una decisión en particular. Y en sentido similar, apunta PALMA ORTIGOSA, A., «Decisiones automatizadas en el RGPD. El uso de algoritmos en el contexto de la protección de datos», *Revista General de Derecho Administrativo*, n. 50, 2019, p. 14/25, que «el derecho a la explicación exige del responsable una aportación de información mucho más personalizada sobre la decisión algorítmica que aquella que debía aportar en un primer momento el responsable conforme a los Arts. 13.2.f), 14.2.g), 15.1.h) RGPD, información previa que no tenía todavía en cuenta la decisión automatizada del particular basada en los datos personales aportados por este último».

gos discriminatorios de un sistema de IA se producirá siempre *ex post*, es decir, una vez adoptada la decisión discriminatoria⁴⁴⁷.

Para un sector doctrinal, al que nos adherimos, se podría defender la existencia de este derecho a la explicabilidad *en ambos momentos*. Una interpretación sistemática de los derechos de información de los arts. 13, 14 y 15, junto con los derechos del art. 22.3 y el Considerando 71, implicaría reconocer que el derecho a la explicación está incorporado siquiera implícitamente en el art. 22.3. Así lo mantiene, entre otros, NUÑEZ SEOANE para quien no cabe duda de que se reconoce este derecho en ambas sedes teniendo en cuenta que es el resultado o consecuencia del ejercicio por el interesado de los previos derechos o garantías a la intervención humana y a formular las alegaciones que prevé el artículo 22.3; así, una vez solicitada la intervención humana y formuladas las alegaciones pertinentes, el responsable deberá ofrecer «una explicación de la decisión tomada después de tal evaluación» a la que se refiere el Considerando 71, ya que solo *después* de que sus alegaciones hayan sido revisadas, evaluadas y explicadas por una persona, podrá ejercer en condiciones su derecho de impugnación⁴⁴⁸.

Más allá de estas vacilaciones acerca de la propia *existencia* del derecho de explicabilidad, también se ha discutido (y criticado) su *alcance* y *contenido*. Determinar cuándo opera este derecho obliga a analizar con más detalle el art. 22 RGPD por cuanto los derechos y garantías mencionadas se aplican sólo en los casos de decisiones individuales automatizadas y perfilados a que éste se refiere.

⁴⁴⁷ LÓPEZ-TARRUELLA MARTÍNEZ, A., *op.cit.*, p. 197.

⁴⁴⁸ NUÑEZ SEOANE, J., «El derecho a la información y acceso al funcionamiento de los algoritmos que tratan datos personales», en *La regulación de los algoritmos*, coord. por Díaz González y Huergo Lora (dir.) 2020, LÓPEZ-TARRUELLA MARTÍNEZ, A., *op.cit.* p. 308. En el mismo sentido, SELBST, A. D./POWLES, J., «Meaningful information and the right to explanation», *International Data Privacy Law*, Vol. 7, n.º 4, 2017, pp.233 y ss; GOODMAN., B/FLAXMAN, S., «European Union regulations on algorithmic decision-making and a “right to explanation”» presented at ICML Workshop on Human Interpretability in Machine Learning, New York, 2016, <https://arxiv.org/abs/1606.08813v3>; BIBAL A./LOGNOUL, M./DE STREEL, A./FRÉNAY, B., «Legal requirements on explainability in machine learning», *Artificial Intelligence & Law*, 29 (2), 2021, pp. 149-169. También entre nuestros autores, NAVAS NAVARRO, S., *Daños ocasionados ...*, cit., p. 72.

En sentido contrario, la falta de refrendo legal y el carácter no vinculante de los considerandos, ha llevado a otro sector doctrinal a entender que en el RGPD no existe un derecho de explicabilidad. Manteniendo esta última postura, WACHTER, S., MITTELSTADT, B., y FLORIDI, L., «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation», *International Data Privacy Law*, Vol. 7, n. 2, 2017. Estos autores afirman que la omisión del derecho a la explicación en el artículo 22.3 es intencionada, por lo que niegan su reconocimiento. También, en nuestra doctrina, GIL MEMBRADO, C., «Daños producidos por la IA...», cit., p. 521, considera que el RGPD no reconoce el derecho a la explicabilidad *ex post*.

2.5.2.1. *Exégesis del art.22 RGPD*

El art. 22 RGPD es el punto de enlace entre la IA y la protección de datos puesto que las decisiones automatizadas y la elaboración de perfiles son dos de sus aplicaciones más frecuentes⁴⁴⁹. Es más, la *Estrategia Europea sobre IA* se encuentra directamente inspirada por la regulación del RGPD relativa a las decisiones automatizadas.

Se trata de uno de los pocos preceptos que prevé una serie de garantías jurídicas en los casos en que se adopten decisiones automatizadas, ya sea mediante sistemas de IA o no. Por tanto, si el tratamiento de datos personales, la elaboración de perfiles o la toma de decisiones sobre una persona física es realizada por un componente IA, tendrá que someterse al RGPD. En caso contrario, esto es, si no se tratan datos personales, no será necesario. El problema reside en que en muchos casos no es sencillo determinar si durante una etapa del ciclo de vida de un sistema basado en IA se tratan o no datos personales⁴⁵⁰.

Tanto el ámbito de aplicación del precepto, como los derechos y garantías que implica y su concreta efectividad cuando la automatización de decisiones se lleva a cabo a través de sistemas de IA, han sido objeto de críticas y comentarios por la doctrina⁴⁵¹. Además, en fechas muy recientes el TJUE se ha pronunciado por primera vez sobre el alcance y ámbito de aplicación del art. 22 RGPD, en su sentencia de 7 de diciembre de 2023, asunto C634/21 (*OQ y Land Hessen, con intervención de SCHUFA Holding AG*), sentando doctrina

⁴⁴⁹ HERRERA DE LAS HERAS, R., «Protección de datos e inteligencia artificial», en Cruz Blanca/Lledó Benito (coords.), *La robótica y la inteligencia artificial en la nueva era de la revolución industrial 4.0*, 2021, p. 654. No obstante, como matiza REBOLLO DELGADO, L., *Inteligencia artificial y derechos fundamentales*, Madrid, 2023, p. 107, estas dos funcionalidades hoy en día ya no son troncales de la IA y la aplicación del precepto únicamente a ellas se manifiesta claramente insuficiente teniendo en cuenta la amplitud de posibilidades que ofrece la IA. Para una explicación detallada sobre las diferencias entre estas dos figuras jurídicas recogidas en el RGPD que guardan relación con los procesos de tratamiento de datos y que dan lugar a inferencias o predicciones sobre comportamientos o preferencias de las personas, puede verse LAZCOZ MORATINOS, G., *op.cit.*, p. 288. Baste mencionar aquí que las decisiones automatizadas tienen un ámbito de aplicación que puede solaparse parcialmente con la elaboración de perfiles o derivar de ésta. En ocasiones, la toma de decisiones automatizada requiere previamente la elaboración del perfil de la persona intentando hacer una predicción sobre sus características o comportamiento; en otros casos, la decisión se toma sobre la base de datos referentes a la actividad de la persona sin necesidad de elaborar un perfil.

⁴⁵⁰ Así lo afirma la AEPD, *Adecuación...*, cit., p.14.

⁴⁵¹ Por todos, PALMA ORTIGOSA, A., «Decisiones automatizadas en el RGPD...», cit., y REBOLLO DELGADO, L., *op.cit.*, pp. 101 y ss. Para OTTOLIA, A., *Derecho, Biga Bata e Inteligencia Artificial*, Valencia, 2018, p.102, «[...] la aplicación real del precepto es decepcionante, ya que resulta debilitado por las excepciones que prevé».

importante al respecto. Es conveniente, por ello, detenerse en el análisis de tales cuestiones.

A) *Ámbito de aplicación del art. 22 RGPD: alcance de la prohibición y excepciones*

La primera garantía que recoge dicho precepto consiste, precisamente, en la *prohibición general de que las personas puedan ser objeto de decisiones basadas únicamente en el tratamiento automatizado de sus datos personales*, incluyendo la elaboración de perfiles, *siempre que dichas decisiones produzcan efectos jurídicos sobre ellas o les afecten significativamente de modo similar* (art. 22.1).

No obstante, se establecen una serie de casos en los que excepcionalmente se permiten dichas decisiones, a saber: a) cuando sean necesarias para la celebración o la ejecución de un contrato entre el interesado y un responsable del tratamiento; b) cuando estén autorizadas por el Derecho de la Unión o de los Estados miembros que se aplique al responsable del tratamiento y que establezca asimismo medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado; y c) cuando el interesado haya dado su consentimiento explícito.

a) *Excepciones*

La primera excepción plantea muchas dudas interpretativas acerca de cuándo «es necesaria» una decisión de este tipo para la celebración o ejecución de un contrato, ya que lo normal es que el uso de sistemas para automatizar decisiones se lleve a cabo más por conveniencia de una de las partes que por necesidad⁴⁵². Además, la «necesidad» debe interpretarse de manera estricta, lo que significa que el tratamiento automatizado debe ser esencial para poder perfeccionar el acuerdo o cumplir con las obligaciones contractuales. Así, esta base de legitimación podrá alegarse en sectores u organizaciones donde es habitual realizar una evaluación previa del interesado que pretende formalizar un contrato con el responsable. Piénsese por ejemplo en sectores como el bancario, el de seguros o el laboral. De hecho, en el caso de la concesión de créditos es obligatorio realizar una evaluación previa de la solvencia patrimonial del solicitante⁴⁵³. Para ello, el

⁴⁵² HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, p. 62. De hecho, el GT 29 reconoce en sus *Directrices sobre decisiones individuales automatizadas*, *cit.*, p. 26, que los responsables del tratamiento pueden, con fines contractuales, acudir a los procesos de decisiones basadas únicamente en el tratamiento automatizado de datos por considerar que es la mejor manera de conseguir un determinado objetivo, ya que «la participación humana rutinaria puede resultar poco práctica o imposible debido a la magnitud de los datos tratados».

⁴⁵³ Vid., art. 18 de la Orden EHA/2899/2011, de 28 de octubre, de transparencia y protección del cliente de servicios bancarios y art 29 de la Ley 2/2011, de 4 de marzo, de Economía Sostenible. Sobre la obligación de los bancos de evaluar la solvencia del cliente antes de la concesión de un crédi-

responsable, antes de concertar dicho contrato, valorará al interesado con el objetivo de analizar la viabilidad futura del acuerdo. En definitiva, lo realmente importante para que se pueda aplicar esta base legitimadora del tratamiento es que el responsable demuestre que no hay otro método igualmente efectivo y menos invasivo para la intimidad con el que se pueda lograr el mismo objetivo. Si lo hubiera, dicho tratamiento no podría calificarse de «necesario»⁴⁵⁴.

Respecto a la segunda excepción, parece que se permiten las decisiones plenamente automatizadas cuando estén autorizadas legalmente, es decir, se permite que el Derecho de la UE o de los Estados miembros pueda establecer normas donde se autorice a los responsables a llevar a cabo este tratamiento automatizado de datos personales para la toma de decisiones. Del propio tenor de esta disposición se desprende que el Derecho nacional que autorice la adopción de una decisión individual automatizada debe establecer medidas adecuadas para salvaguardar los derechos, libertades e intereses legítimos del interesado⁴⁵⁵. Además, el tratamiento deberá respetar los requisitos establecidos en los arts. 5 y 6 RGPD⁴⁵⁶.

Respecto a la tercera excepción, el consentimiento del interesado como base legitimadora para el tratamiento de datos personales es la clave de bóveda de la normativa de protección de datos. Sin embargo, no garantiza en absoluto un conocimiento y comprensión suficientes de las consecuencias que implica la cesión y el tratamiento de dichos datos. Lo mismo ocurre con las decisiones automatizadas puesto que los interesados no son conscientes de la trascendencia e impacto de sus resultados ni de los posibles sesgos discriminatorios que pueden esconder. En cualquier caso, este consentimiento debe cumplir con los requisitos del art. 4.11 en relación con el Considerando 42 RGPD, es decir, debe ser informado, específico, inequívoco, libre y revocable⁴⁵⁷.

to, puede verse MARÍN LÓPEZ, M.J., «La obligación de evaluar la solvencia del prestatario en la Ley 5/2019, de contratos de crédito inmobiliario», *Revista de Derecho Patrimonial*, n. 50/2019, accesible en https://manueljesusmarin.es/wp-content/uploads/2022/02/doctrin_084.pdf

⁴⁵⁴ Así lo explica el GT29 en sus *Directrices sobre decisiones individuales automatizadas*, cit., p.26.

⁴⁵⁵ El Considerando 71 RGPD establece que las medidas a adoptar deben incluir, en particular, la obligación del responsable del tratamiento de utilizar procedimientos matemáticos o estadísticos adecuados, de aplicar las medidas técnicas y organizativas apropiadas para garantizar que se reduzca al máximo el riesgo de error y se corrijan errores, y de asegurar los datos personales de forma que se tengan en cuenta los posibles riesgos para los intereses y derechos del interesado y se impidan, entre otras cosas, los efectos discriminatorios en las personas físicas.

⁴⁵⁶ Como indica el TJUE, en su sentencia de 7 de diciembre de 2023, párrafos 65, 66 y 68, los Estados miembros no podrán adoptar normativas que autoricen las decisiones automatizadas o elaboración de perfiles incumpliendo los requisitos establecidos en los arts. 22.2.b) y 22.4 y en los arts. 5 y 6 RGPD. Para un mayor desarrollo de esta excepción puede verse PALMA ORTIGOSA, A., «Decisiones automatizadas en el RGPD...», cit., p. 9/25.

⁴⁵⁷ Analiza cada uno de estos requisitos, CASADESÚS RIPOLL, P., *op.cit.*, pp. 149-154.

En tales supuestos excepcionales, y más concretamente en los previstos en las letras a) y c), el precepto recoge una serie de garantías «mínimas» que deberá adoptar el responsable para salvaguardar los derechos e intereses legítimos del afectado por la decisión automatizada: concretamente, el *derecho a obtener intervención humana* por parte del responsable, a *expresar su punto de vista* (opinar sobre el tratamiento) y a *impugnar*⁴⁵⁸ la decisión (art.22.3). A las que se debe añadir, tal como menciona el Considerando 71, la *información específica* al interesado. Las mismas garantías, al menos, deberá adoptar el Estado que autorice este tipo de decisiones en su ordenamiento en virtud del art. 22.2.b).

A continuación, se establecen dos importantes precisiones respecto a este tipo de decisiones que deben tenerse en consideración: por un lado, que la decisión automatizada no puede basarse en las categorías de datos especialmente protegidos del art. 9.1, salvo excepciones entre las que se incluyen «razones de un interés público esencial» (art. 22.4); y por otro lado, que mediante Ley se podrá limitar el alcance de las obligaciones y derechos establecidos en el art. 22 en los numerosos supuestos del art. 23 (art.23.1).

b) *La prohibición general del art. 22 RGPD*

Dos son los presupuestos que exige el art. 22.1 para que opere la prohibición que contiene: por un lado, que se trate de decisiones basadas únicamente en el tratamiento automatizado de datos personales, y por otro, que dichas decisiones tengan *efectos jurídicos, o significativamente similares*, en el interesado.

1. *Primer presupuesto: que se trate de una decisión exclusivamente automatizada*

El primer presupuesto implica el uso de sistemas que de forma automática y sin intervención humana significativa adopten una decisión que afecte a la persona (por ejemplo, decidir qué candidatos para un puesto de trabajo se descartan en un proceso de selección y quiénes pasan a la fase de entrevista, rechazar automáticamente la concesión de un crédito o la concertación de un seguro, etc). Como se encargó de aclarar el GT29, para que se trate de decisiones exclusivamente automatizadas no debe haber una intervención humana responsable que analice la información procesada y cuente con poder o capacidad de modificar la decisión⁴⁵⁹.

⁴⁵⁸ Sobre este concreto derecho, RIVAS VALLEJO, P. «Análisis...», *cit.*, p. 460, dice que «[t]ampoco se matiza si el derecho a impugnar la decisión tiene un marco de acción reducido a la interacción con “el responsable del tratamiento” o si tiene un contenido más amplio equivalente al derecho a reclamar contra la decisión empresarial, que debería entenderse en sentido interno, pues la legislación garantiza en todo caso el derecho a la reclamación administrativa o judicial (art. 24 CE)».

⁴⁵⁹ *Directrices sobre decisiones individuales automatizadas*, *cit.*, p. 23. Sobre lo que implica esta exigencia puede verse también RIVAS VALLEJO, P. «Análisis...», *cit.*, p. 457 y ss.

Surge la duda de si la prohibición se aplicaría a los casos en que esa intervención humana relevante se limitara a una o varias partes del proceso de decisión, pero no a todas. Podría defenderse que, en la medida en que dicha intervención humana significativa exista (aunque sea en una etapa del proceso), las decisiones no quedarían bajo el ámbito de aplicación del art. 22. Por tanto, en tales casos el afectado no gozaría de los derechos y garantías anteriormente comentadas. No obstante, creemos más acertado mantener que para poder eludir la aplicación del precepto, la existencia de intervención humana debería apreciarse en cada una de las decisiones parceladas dirigidas al mismo fin y no solamente en alguna parte del proceso. Así, por ejemplo, si un algoritmo selecciona las 10 mejores candidatas a un puesto de trabajo de entre las 1000 que se postulan, descartando automáticamente a las demás, existiría una decisión automatizada respecto de la que habría que garantizar los derechos recogidos en la norma, aunque las 10 últimas fueran evaluadas por una persona⁴⁶⁰. Por el contrario, el art. 22 sería inaplicable si en ese proceso ha intervenido un supervisor que ha decidido quién pasa a la segunda fase y quién no, sobre la base de la recomendación que hace el sistema IA de apoyo a la decisión humana (y que, si lo considera oportuno, puede no seguir), salvo que esa intervención humana hubiera sido exclusivamente simbólica o ficticia, asumiendo sin más lo decidido por el sistema⁴⁶¹.

En definitiva, para eludir tanto la prohibición como las obligaciones que impone el art. 22 a los responsables del tratamiento cuando hay base legítima del tratamiento automatizado, debe quedar claro que la intervención humana debe ser significativa en el sentido de ser realizada por una persona con competencia y autoridad sobre la decisión y que valore toda la información disponible, no siendo suficiente con que se limite a convalidar, replicar o reproducir acríticamente la decisión o recomendación del sistema. Su inter-

⁴⁶⁰ MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL, *Información algorítmica en el ámbito laboral. Guía Práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral*, mayo 2022, p.10. Accesible en file:///C:/Users/ASUS/Downloads/algortca_laboral-I.pdf. Como explica RIVAS VALLEJO, P. «Análisis...», *cit.*, p. 456, «la norma parece apuntar a un ámbito restringido del citado derecho a la explicabilidad, no extensible, por tanto, a decisiones automatizadas en parte del proceso de decisión, aun cuando esta parte sea especialmente relevante en el conjunto de factores que conducen a la decisión final». El problema de mantener esta interpretación, como sigue diciendo la autora, es que «si se considera que han sido presentados mil currículos, de los cuales han sido preseleccionados quince para su examen humano, el grueso del sesgo, donde probablemente radique la mayoría de discriminaciones por diversas causas protegidas, habrá quedado incluido en una fase íntegramente automatizada, sobre la cual no se ofrecerá explicación alguna».

⁴⁶¹ Como mantiene EGUÍLUZ CASTAÑEIRA, J.U., *op.cit.*, p. 344, si no se entendiera que esa intervención humana no ha de limitarse a reproducir la decisión del sistema, las empresas podrían eludir fácilmente la aplicación del art. 22, destinando un empleado a supervisar decisiones automatizadas, que en la práctica se limitara a asumir acríticamente sus resultados.

vención no debe consistir en una mera supervisión superficial de los resultados del algoritmo sino en una participación real, relevante o sustancial en el proceso decisorio.

Del análisis de este primer presupuesto puede extraerse, como conclusión, que la regulación no afecta a todas las predicciones algorítmicas sino solo a aquellas que se integran en sistemas de decisión plenamente automatizados⁴⁶². Por ello, y sobre la base de que en la práctica lo normal es que las decisiones no suelen tomarse de forma plena o puramente automática (sobre todo cuando son de una entidad considerable), sino como un elemento más de juicio, junto con otros, para tomar la decisión final, el art. 22 RGPD perdería cierta relevancia puesto que se aplica exclusivamente a una de las distintas formas de utilización de estos sistemas algorítmicos⁴⁶³. Es decir, la prohibición que contiene gozaría de escasa aplicación en tanto en cuanto no abarca las decisiones semiautomatizadas y, en la actualidad, no es tan frecuente la toma de decisiones plenamente automatizadas que cumplan las condiciones que exige la norma. Lo más normal es que se utilicen los resultados de estos sistemas como elemento auxiliar que el operador humano puede tener en cuenta, o no, para adoptar su decisión final⁴⁶⁴.

Dicha limitación supone una importante traba en la lucha contra la opacidad de los sistemas de IA (y por lo que a nosotros interesa, contra la discriminación algorítmica) por cuanto los sistemas que se usan como apoyo a la decisión también pueden arrojar resultados incorrectos y discriminatorios al contener, por ejemplo, sesgos en el diseño o en los conjuntos de datos utilizados⁴⁶⁵, y principalmente, por el denominado «sesgo de la automatización», es decir, la tendencia comprobada de que los humanos confiamos o creemos (erróneamente) que el resultado ofrecido por la máquina es mucho más acertado, objetivo y neutral que el nuestro, cuando ambos juicios no coinciden⁴⁶⁶.

Por ello, en aras de una mejor defensa de los intereses del perjudicado por la decisión, señala el GT29 que, aunque la decisión automatizada o la elaboración de perfiles no cumplan la definición del art. 22.1, seguiría siendo aconsejable ofrecer la información ampliada o reforzada prevista en el

⁴⁶² En este sentido, por todos, HUERGO LORA, A., «Una aproximación...», cit., p. 60 o GINÈS i FABRELLAS, A., *op.cit.*, p. 319. También se hace referencia a esta idea en MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL, *Información algorítmica en el ámbito laboral...*, cit., p. 15, cuando se dice que la regulación europea usa un concepto restrictivo de algoritmo en la medida en que exige que tome decisiones íntegramente automatizadas.

⁴⁶³ PALMA ORTIGOSA, A., «Decisiones automatizadas en el RGPD...», cit., p. 5/25.

⁴⁶⁴ Por todos, HUERGO LORA, A., «Una aproximación...», cit., p. 71.

⁴⁶⁵ GINÈS i FABRELLAS, A., *op.cit.*, p. 319.

⁴⁶⁶ Nos remitimos a lo ya dicho sobre este sesgo en el Capítulo I, apartado 2.2.1.

RGPD⁴⁶⁷. Ahora bien, a nuestro modo de ver, más allá de esta recomendación o consejo, lo que habría que plantearse sería la expresa ampliación del ámbito de aplicación de este derecho de información específico y demás garantías previstas en el precepto a los casos de decisiones semiautomatizadas, tal y como por ejemplo ha hecho el legislador laboral en el art. 64.4 ET⁴⁶⁸ y como parece sostener también el TJUE en su sentencia de 7 de diciembre de 2023 que será analizada *infra*.

2. Segundo presupuesto: efectos jurídicos o similares

Respecto al segundo presupuesto exigido por el art. 22 RGPD, esto es, que la decisión automatizada produzca *efectos jurídicos o significativamente similares* en las personas, se trataría de que la decisión afectara de alguna manera a los derechos o la esfera jurídica de la persona, como por ejemplo sucede cuando te deniegan un préstamo o una subvención pública, te excluyen de un proceso de selección laboral o te despiden de un trabajo⁴⁶⁹. Es esencial,

⁴⁶⁷ Vid., *Directrices sobre decisiones individuales automatizadas*, cit., p. 28.

⁴⁶⁸ Esta reforma del ET fue realizada a través del Real Decreto-ley 9/2021, de 11 de mayo, para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales, que fue convalidado posteriormente por la Ley 12/2021, de 28 de septiembre (conocida como *Ley rider*). Se modifica el art. 64.4 ET para incluir un derecho de información colectiva, que se suma al derecho de información individual previsto en el RGPD, y que se aplica aun cuando se trate de *sistemas semiautomatizados* de toma de decisión. En virtud de este derecho, la representación legal de la plantilla tendrá que ser informada de «los parámetros, reglas e instrucciones en los que se basan los algoritmos o sistemas de inteligencia artificial que afectan a la toma de decisiones que puedan incidir en las condiciones de trabajo, el acceso y mantenimiento del empleo, incluida la elaboración de perfiles». Sobre este derecho señala GINÈS i FABRELLAS, A., *op.cit.*, p. 319 que, si bien se trata de una regulación pionera a nivel europeo que permite a la representación legal de la plantilla valorar la legalidad –e, incluso, potencial discriminatorio– de las variables utilizadas por el algoritmo, debería haberse acompañado con información estadística sobre el impacto del uso de dicha tecnología inteligente.

⁴⁶⁹ Pueden verse otros ejemplos de este tipo de decisiones en las *Directrices sobre decisiones individuales automatizadas*, cit., pp. 23-24. También en el Considerando 71 RGPD se pone como ejemplo la denegación automática de una solicitud de crédito en línea o los servicios de contratación en red en los que no medie intervención humana alguna. Sin embargo, a pesar de dichos ejemplos, no resulta siempre evidente cuándo una decisión afecta significativamente a un sujeto. Como afirman MENDOZA, I./ BYGRAVE, L.A. «The Right Not to Be Subject to Automated Decisions Based on Profiling», *University of Oslo Faculty of Law Research Paper* n.20, 2017, p. 12, aunque puede deducirse que en aquellos casos en los que la decisión tenga impacto en la situación financiera o laboral del interesado estaremos ante un efecto significativo, el efecto concreto dependerá de las características de cada individuo, por lo que se deberá hacer un estudio *ad casum* sobre la significación o relevancia del efecto. Para EGUÍLUZ CASTANEIRA, J.U., *op.cit.*, pp. 342 y ss., un ejemplo de una decisión con «efectos jurídicos» sería una decisión judicial, o una decisión con respecto a un beneficio social otorgado por la ley, como el pago de la pensión. Por otro lado, un ejemplo de una decisión con efectos «significativamente similares» sería un banco que niega un crédito automáticamente. Y en base a lo que sostienen las Autoridades de Protección de Datos, la diferenciación de precios en línea podría

por tanto, que la decisión tenga el potencial de afectar significativamente a las circunstancias, al comportamiento o a las elecciones de las personas afectadas, impacte de forma prolongada o permanente en el interesado, o, en los casos más extremos, provoque la exclusión o discriminación de la persona⁴⁷⁰.

Algún autor pone en duda que esta clase de decisiones (que a veces ni siquiera se refieren a personas concretas, sino a grupos o a zonas) produzcan los «efectos jurídicos en el interesado» a que se refiere el precepto⁴⁷¹. Sin embargo, incluso en la conocida sentencia del caso *SyRI* del Tribunal de Distrito de La Haya, de 5 de febrero de 2020, se reconoce que los informes de riesgo elaborados por el sistema de IA acerca de la probabilidad de que determinados ciudadanos defraudaran al fisco, les afectaban «significativamente» de un modo similar a una decisión con efectos jurídicos.

Las decisiones automatizadas basadas en datos personales que no tengan efectos o consecuencias jurídicas para los sujetos, no quedarían bajo el ámbito de aplicación del precepto analizado. Como ejemplo de este tipo de decisiones HERRERA DE LAS HERAS pone el de un restaurante que utiliza un sistema de reconocimiento facial y detección de emociones para determinar el grado de satisfacción del cliente. La decisión del sistema acerca del grado de satisfacción concreto de cada comensal no tendría efectos jurídicos para ellos en la medida en que no supondría que se les denegara el acceso en próximas visitas, o que se les cobrara más o menos en función de su grado de satisfacción, sino que simplemente iría dirigida a evaluar su grado de satisfacción. En opinión del autor citado, parece que en estos casos el cliente no podría negarse al escrutinio hecho por el sistema inteligente⁴⁷². No obstante, a nuestro modo de ver, si bien no tendría los derechos y garantías del art. 22.3 ni el derecho de información reforzado de los arts. 13.2.f), 14.2.g) y 15.1.h), sí tendría derecho a ser informado del resto de extremos que contemplan tales normas y gozaría del resto de derechos que otorga la normativa cuando se procede al tratamiento de datos personales (acceso, rectificación, oposición...).

Otro ejemplo de decisiones automatizadas basadas en datos personales que no tendrían efectos o consecuencias jurídicas para la persona y, por lo

«afectar significativamente de manera similar» a alguien, si conduce a «precios prohibitivamente altos que efectivamente excluyen a alguien de ciertos bienes o servicios».

⁴⁷⁰ SÁEZ LARA, C., *op.cit.*, p. 53.

⁴⁷¹ Así, HUERGO LORA, A., «Una aproximación...», *cit.*, p. 61.

⁴⁷² HERRERA DE LAS HERAS, R., «Protección de datos e inteligencia artificial», *cit.*, p. 655.

tanto, no quedarían bajo el paraguas del art. 22 RGPD, podría ser el de la personalización de contenido en plataformas de *streaming*. Plataformas como Netflix, Spotify o YouTube utilizan algoritmos automatizados para analizar los datos de comportamiento de los usuarios, como las películas, series, canciones, o videos que han visto o escuchado, las calificaciones que han dado y el tiempo que pasan consumiendo cierto tipo de contenido. Basándose en estos datos, la plataforma genera recomendaciones personalizadas de contenido que pueden interesar a cada usuario (esta sería la decisión automatizada, en este caso en forma de recomendación). Esa recomendación no tendría efectos jurídicos o similares en el usuario porque no alteraría en modo alguno sus derechos u obligaciones, ni influiría en sus oportunidades laborales, su situación financiera, acceso a bienes o servicios esenciales, ni a su reputación, de una manera que pudiera considerarse significativa. Además, la persona no estaría obligada a seguir dichas recomendaciones. De hecho, podría ignorarlas completamente y buscar contenido por sí misma, sin perjuicio de que también pudiese desactivarlas o ajustarlas, controlando así el proceso de personalización. Por último, el impacto que podría tener en la persona esta decisión plenamente automatizada sería trivial y se limitaría a mejorar su experiencia facilitando la búsqueda de contenido que le pudiera interesar.

Por el contrario, un ejemplo de las decisiones contempladas en el art. 22.1 que, en los casos excepcionales en que están permitidas otorgarían una serie de derechos al interesado, podría ser el rechazo automático de una solicitud de préstamo bancario basado únicamente en el perfil crediticio del solicitante generado por un algoritmo. Este tipo de decisión sí que tendría un efecto jurídico directo sobre la persona porque afectaría a su capacidad para obtener un crédito y, potencialmente, a su situación financiera general.

Presentes los dos presupuestos anteriores serían aplicables las garantías del art. 22.3 RGPD así como la regulación específica en relación al derecho de información y acceso del interesado previsto en los arts. 13.2 f), 14.2 g) y 15.1 h) ⁴⁷³, en cuya virtud, la persona afectada por la decisión tendrá *derecho a saber que está siendo objeto de una decisión automatizada*, es decir, derecho a conocer la existencia en sí del tratamiento automatizado de datos que le afectará, así como el derecho a que se le proporcione *información significativa sobre la lógica aplicada por el sistema y sobre la importancia y las consecuencias previstas de dicho tratamiento*. En consecuencia, los responsables estarían obligados a informar de ello (así como de su base legitimadora dada la prohibición general del art.22.1) y, al menos en tales casos, de los extremos anteriores. Por el contra-

⁴⁷³ Como explican las *Directrices sobre decisiones individuales automatizadas*, cit., del GT29, p. 28, «ofrecer esta información también ayudará a los responsables del tratamiento a garantizar que cumplan algunas de las garantías obligatorias descritas en el artículo 22, apartado 3, y el considerando 71».

rio, no se aplicarían tales derechos y garantías si las decisiones no cumplieran los requisitos del art. 22, bien porque hubiera existido intervención humana significativa en su adopción (casos de decisiones semiautomatizadas), bien porque no desplegaran efectos jurídicos o similares sobre las personas.

Si bien esta conclusión era la que se venía manteniendo hasta el momento, la reciente sentencia del TJUE, de 7 de diciembre de 2023, ha sentado una interpretación distinta por lo que se refiere a los presupuestos del art. 22 RGPD, y en concreto, a su aplicación a las decisiones semiautomatizadas. Veámosla.

3. *La STJUE de 7 de diciembre de 2023, asunto C634/21 (OQ y Land Hessen, con intervención de SCHUFA Holding AG)*

Precisamente en la línea apuntada parece situarse la reciente sentencia del TJUE, de 7 de diciembre de 2023⁴⁷⁴, donde el tribunal realiza una interpretación extensiva del art. 22 RGPD a fin de aplicar las garantías que contiene también a las decisiones parcial o semiautomatizadas⁴⁷⁵. El caso versaba sobre un litigio entre SCHUFA, una empresa alemana que proporciona información de solvencia a terceros, y OQ, un particular a quien un banco le denegó un préstamo en base a la información negativa que SCHUFA proporcionó sobre él. OQ ejerció su derecho de acceso a los datos registrados y su eliminación, pero SCHUFA se negó a revelar todos los datos utilizados en su evaluación. Solo cedió información genérica pero no los datos que tenía del afectado ni la ponderación de los mismos que había realizado en forma de valores de probabilidad. Justificó su negativa en que no había sido ella quien

⁴⁷⁴ Sentencia del Tribunal de Justicia de la Unión Europea (Sala Primera). Asunto C-634/21 (SCHUFA Holding (Scoring)) y en los asuntos acumulados C-26/22 y C-64/22 (SCHUFA Holding (Exoneración del pasivo insatisfecho)) de 7 de diciembre de 2023.

⁴⁷⁵ Sobre esta resolución puede verse, COTINO HUESO, L., «La primera sentencia del Tribunal de Justicia de la Unión Europea sobre decisiones automatizadas y sus implicaciones para la protección de datos y el Reglamento de inteligencia artificial» Diario LA LEY, N° 80, Sección Ciberderecho, 17 de Enero de 2024. Recuperado de La Ley Digital, LA LEY 1301/2024. Para este autor «es reconfortante observar un progresivo reconocimiento normativo de la importancia de las garantías en decisiones parcial o semi-automatizadas», teniendo en cuenta que tanto la Carta de Derechos Digitales en España, como algunas legislaciones de protección de datos fuera de la UE también las recogen expresamente, como por ejemplo, la de Ecuador. También en Canadá o EEUU la definición de sistema de decisiones automatizado (*automated decision system*) incluye tanto las decisiones totalmente automatizadas como las de apoyo a la decisión. Vid., por ejemplo, el proyecto de Ley de Responsabilidad Algorítmica de 2022 de EEUU (sección 2; Definiciones).

había denegado el crédito sino sus socios contractuales (es decir, el banco) e invocó también secretos comerciales⁴⁷⁶.

El Tribunal de lo Contencioso-Administrativo de Wiesbaden planteó dos cuestiones prejudiciales al Tribunal de Justicia sobre la interpretación del art. 6 y 22.1 RGPD, con el fin de determinar, por un lado, si la actividad realizada por SCHUFA (esto es, la generación automatizada de un valor de probabilidad transmitido a terceros que toman decisiones basadas en él) constituía una «decisión automatizada» en el sentido del art. 22 RGPD teniendo en cuenta que no era SCHUFA quien tomaba la decisión final, y por otro lado, si la negativa de SCHUFA a proporcionar información adicional al afectado suponía una vulneración de los derechos de protección de datos de OQ por estar obligada a ello, teniendo en cuenta que los terceros que tomaron la decisión final –en base a dicho valor de probabilidad– no disponían de tal información.

En su sentencia, el Tribunal se detiene en el análisis de los presupuestos de aplicación del art. 22, que concreta en tres: en primer lugar, la existencia de una «decisión»; en segundo lugar, la necesidad de que ésta se base «únicamente» en un tratamiento automatizado de datos personales (incluida la elaboración de perfiles); y en tercer lugar, que la decisión despliegue «efectos jurídicos» en el interesado o le «afecte significativamente de modo similar».

Respecto al concepto de «decisión» al que se refiere el art. 22.1 RGPD, sostiene una interpretación extensiva en la línea mantenida por el Abogado General en el punto 38 de sus conclusiones. Entiende que puede incluir diversos actos con potencial para afectar al interesado de múltiples maneras y que engloba el resultado del cálculo de la solvencia de una persona en forma de un valor de probabilidad relativo a su capacidad para hacer frente a sus compromisos de pago en el futuro (párrafo 46).

En relación al requisito de que la decisión deba estar «basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles», sostiene, como señalaba también el Abogado General en el punto 33 de sus conclusiones, que no se discute que una actividad como la que realiza SCHUFA responde a la definición de «elaboración de perfiles» del art.4.4 del RGPD y, por tanto, que en el presente asunto se cumple este requisito, habida cuenta de que el tenor de la primera cuestión prejudicial se refiere expresamente

⁴⁷⁶ Las entidades de información y análisis de riesgo (como SCHUFA o en el caso del mercado español, Equifax o Experian) analizan el riesgo de un prestatario, combinando factores objetivos y subjetivos a partir de la información de que disponen sobre el prestatario, y/o de la información que sus socios comerciales (por ejemplo, los bancos) les proporcionan con el fin de que calculen esa puntuación o valor de probabilidad determinado. No obstante, son dichos socios comerciales los que toman la decisión final sobre la ejecución (o modo de ejecución) de un contrato con el prestatario, es decir, los que deciden la concesión o denegación del crédito, si bien con base en dicho valor.

a la generación automatizada de un valor de probabilidad a partir de datos personales relativos a una persona y acerca de su capacidad para satisfacer un préstamo en el futuro (párrafo 47).

Por lo que se refiere al requisito de que la decisión deba producir «efectos jurídicos» en el interesado o «[afectarlo] significativamente de modo similar», sostiene que del propio tenor de la primera cuestión prejudicial se desprende que la acción del tercero al que se transmite el valor de probabilidad (el banco) está basada «de un modo determinante» en dicho valor. Así pues, según las apreciaciones de hecho del órgano jurisdiccional remitente, en el caso de que un consumidor solicite un préstamo a un banco, un valor de probabilidad insuficiente dará lugar a que el banco deniegue la concesión del préstamo solicitado en la práctica totalidad de los casos (párrafo 48). En estas circunstancias, considera que la decisión sobre el valor de probabilidad, cuanto menos, afecta al interesado significativamente (párrafo 49)⁴⁷⁷.

Por tanto, el TJUE entiende que la generación automatizada del valor de probabilidad proporcionado por SCHUFA y utilizado por el banco para tomar decisiones sobre créditos, debe ser calificada como una decisión individual automatizada (basada en una elaboración de un perfil) que afecta significativamente al interesado. Ello implica la necesidad de cumplir con las disposiciones del RGPD sobre protección de datos y determina la obligación de SCHUFA de divulgar información adicional al interesado⁴⁷⁸.

Esta interpretación se ve corroborada por el contexto en el que se inscribe el art. 22.1 RGPD, así como por los objetivos y la finalidad que persigue dicho Reglamento.

Según mantiene dicho Tribunal, de no mantener esta interpretación existiría una laguna jurídica que mermaría la aplicación, finalidad y protección que se pretende con el art. 22 RGPD. Una interpretación restrictiva del precepto que ex-

⁴⁷⁷ Sobre las dificultades prácticas que pueden derivarse de la consideración de un valor como «determinante» o «no determinante», pueden verse las consideraciones de NEGRO, A., Y SANCHEZ, A., «El veredicto del TJUE sobre el *credit scoring*», accesible en <https://www.cuatrecasas.com/es/spain/propiedad-intelectual/art/veredicto-tjue-sobre-credit-scoring>. Observan, entre otras cosas, que aunque el TJUE nada indica al respecto, cabe entender que la calificación de un valor como «determinante» no depende de que sea el único valor o puntuación que se tenga en cuenta para la toma de las decisiones pertinentes, *sino de que del valor trasladado haya dependido la decisión final* (la cursiva es nuestra).

⁴⁷⁸ Esta conclusión la recoge el párrafo 73 en los términos siguientes: el art.22.1 del RGPD «debe interpretarse en el sentido de que la generación automatizada, por una agencia de información comercial, de un valor de probabilidad a partir de datos personales relativos a una persona y acerca de la capacidad de esta para hacer frente a compromisos de pago en el futuro constituye una “decisión individual automatizada”, en el sentido de la mencionada disposición, cuando de ese valor de probabilidad dependa de manera determinante que un tercero, al que se comunica dicho valor, establezca, ejecute o ponga fin a una relación contractual con esa persona».

cluyese decisiones como la controvertida (esto es, la generación automatizada de un valor de probabilidad) por considerarlas un mero *acto preparatorio* ejecutado por una agencia comercial, calificando únicamente como «decisión» del art. 22 el acto final adoptado por la tercera entidad, generaría un riesgo de elusión del precepto ya que los interesados no podrían ejercitar sus derechos frente a la empresa que calculó su probabilidad de impago, ni siquiera cuando (como ocurrió en este caso) el tercero que hubiera adoptado la decisión final no tuviera toda la información relevante acerca del tratamiento de datos realizado sobre interesado, al haber recibido, únicamente, el valor de probabilidad solicitado⁴⁷⁹.

En definitiva, como señala COTINO HUESO, para el TJUE las garantías del art. 22 RGPD también se pueden aplicar a los actos, hechos o datos sobre los que se sustentan las decisiones automatizadas, como los perfiles o ponderaciones automatizadas que han estado en la base de la decisión finalmente adoptada⁴⁸⁰. Ello a pesar de que *formalmente* la decisión final se adopte por una entidad distinta a la que hace el perfilado o ponderación previa, o de que haya aparentemente una intervención humana. Por ende, esta sentencia expande el alcance del art. 22 RGPD más allá del responsable formal de la decisión, para abarcar a aquellos que procesen datos y que efectúan *materialmente* el tratamiento automatizado para el responsable bajo otras posiciones jurídicas, como la de encargado por cuenta de quien finalmente toma la decisión⁴⁸¹.

⁴⁷⁹ Vid., párrafos 61 y 63.

⁴⁸⁰ Ello supone, por un lado, que el responsable del tratamiento está sujeto a obligaciones adicionales de información en virtud de los artículos 13, apartado 2, letra f), y 14, apartado 2, letra g), de dicho Reglamento. Por otro lado, el interesado tiene derecho, en virtud del artículo 15, apartado 1, letra h), del mencionado Reglamento, a obtener del responsable del tratamiento, en particular, «información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas de dicho tratamiento para el interesado» (párrafo 56). Estos requisitos más estrictos para la licitud de una toma de decisiones automatizada, así como las obligaciones adicionales de información del responsable del tratamiento y los derechos de acceso adicionales del interesado correspondientes se aplican por la finalidad que persigue el artículo 22 del RGPD, que consiste en proteger a las personas contra los riesgos específicos para sus derechos y libertades que supone el tratamiento automatizado de datos personales, incluida la elaboración de perfiles (párrafo 57).

⁴⁸¹ COTINO HUESO, L., «La primera sentencia del Tribunal de Justicia de la Unión Europea...», *cit.*, se refiere a los sujetos intervinientes como «responsable formal de la decisión» (en este caso el Banco) y «tercero encargado del tratamiento de datos» (en este caso, Schufa) y considera que, en base a este pronunciamiento, las «garantías del artículo 22 RGPD se proyectan y van más allá del responsable –al menos formal– que toma la decisión respecto del afectado, generando obligaciones a un tercero encargado que ha tratado datos del afectado». No obstante, para otros autores, las agencias como Schufa que prestan servicios de *credit scoring* actuarán directamente en calidad de «responsables del tratamiento» cuando concurren los tres requisitos analizados por el TJUE y precisamente por ello deberán cumplir las obligaciones que el RGPD impone a los mismos: por un lado, asegurar la aplicación de alguna de las excepciones previstas en el artículo 22.2 y contar con una base legitimadora el tratamiento de los datos de los prestatarios que deberán probar. Asimismo, deberán garantizar el cumplimiento de la normativa aplicable (por ejemplo, aplicando las medidas técnicas

Concluimos, junto al autor citado, que el criterio garantista que adopta el TJUE en esta resolución respecto de las decisiones automatizadas sienta un precedente que podrá proyectarse en muchos otros supuestos en el futuro. Precedente que, además, a nuestro juicio era totalmente necesario para subsanar el insuficiente ámbito de aplicación del art.22, teniendo en cuenta que, al menos en la práctica actual, es mucho más frecuente tomar decisiones con apoyo en los resultados de un sistema de IA, que delegar plenamente en ellos las decisiones finales⁴⁸².

B) *Contenido del derecho de información de los arts. 13.2 g), 14.2 g) y 15.1 h)*

El RGPD no concreta qué implica el deber de proporcionar «*información significativa sobre la lógica aplicada*» y si dicha expresión conlleva o no el derecho a conocer el código fuente del algoritmo y su funcionamiento. Dicha omisión ha hecho que surjan dudas y vacilaciones acerca de esta cuestión⁴⁸³. Es opinión ampliamente compartida que esa disposición del RGPD no im-

y organizativas apropiadas para la seguridad de la información y la reducción del riesgo de error al máximo o garantizar los derechos de los interesados). Así lo mantienen NEGRO, A., Y SANCHEZ, A., *op.cit.loc.cit.* Sea como fuere, en ambos casos los autores coinciden en que los sujetos que *materialmente* efectúan el tratamiento automatizado, perfilado o ponderación de datos quedarían sujetos a ciertas obligaciones.

⁴⁸² No obstante, como augura la doctrina, esta situación parece que se revertirá en un futuro no muy lejano. En este sentido, afirma PALMA ORTIGOSA, A., «Decisiones automatizadas en el RGPD...», *cit.*, p. 6/25, que «el paso del tiempo hará que las decisiones automatizadas que ahora suelen utilizarse como un parámetro más de apoyo a la hora de tomar una decisión determinada, se conviertan en la regla a seguir, pudiendo llegar un momento de estandarización donde se pase de la desconfianza de la máquina, al recelo de la decisión que pueda adoptar un humano contraria a la que ha indicado el algoritmo, de manera que ya no se dude de la máquina sino de la persona».

⁴⁸³ Sobre este asunto planteaba RUBI PUIG dos posibilidades: que la información a suministrar tuviera que consistir en una explicación general del modelo (*model-based explanation*) o en explicar cómo se adoptó una decisión basada en los datos particulares del afectado (*subject-based explanation*). Vid., «Elaboración de perfiles y personalización de ofertas y precios ...», *cit.*, p. 17/24. Sobre esta doble posibilidad, aclara MEDINA GUERRERO, M., *op.cit.*, p. 161, que el hecho de que la información deba facilitarse al recabarse los datos o en un plazo breve de tiempo después (ex arts. 13.2 y 14.3a) hace que en esta fase (previa a la toma de decisión) la obligación de informar no pueda traducirse en la exigencia de asegurar la explicabilidad, ni la fundamentación de una decisión en particular. Lo que debe ponerse en conocimiento del interesado es *la descripción de los tipos de datos y los factores que toma en consideración el sistema, las categorías del árbol de decisión y la ponderación de tales factores a nivel global* (la cursiva es nuestra). Con esa información el afectado debe poder prever qué papel juegan (o pueden jugar) sus diferentes datos objeto de tratamiento en la formación de la decisión a tomar, así como estar en condiciones de acomodar su conducta a los criterios que resultan determinantes para adoptar la decisión. Para NAVAS NAVARRO, S., «La perspectiva de género...», *cit.*, se puede interpretar que la lógica de la que se debe informar en estos casos es la lógica computacional empleada y se trataría de que se pudiera conocer, en la medida de lo posible, el proceso de toma de decisiones para ver si existe algún sesgo o error en el procesamiento de los datos.

pone una obligación de informar y dar acceso de forma exhaustiva al código fuente del algoritmo utilizado para adoptar la decisión, lo que, por otra parte, tampoco sería de utilidad en muchos casos para el interesado si no se comparan también los conjuntos de datos utilizados, por cuanto que en los sistemas adaptativos el código fuente va cambiando⁴⁸⁴. Según las *Directrices sobre decisiones individuales automatizadas* del GT29, se trataría de ofrecer información conforme al *principio de transparencia* en términos sencillos y comprensibles, pero suficientemente exhaustiva para que el interesado pudiera entender *los motivos de la decisión*, sin necesidad de incluir una explicación compleja sobre los algoritmos utilizados, lo que podría ser opaco, confuso, e incluso conducir a la fatiga informativa⁴⁸⁵. A nuestro modo de ver, el *quid* de la cuestión sería que la información proporcionada permitiera al interesado hacer efectivos los derechos de los que es titular. Por tanto, creemos junto con un sector doc-

⁴⁸⁴ Según concluye el GT29 en sus *Directrices sobre decisiones individuales automatizadas*, p. 28, dicho deber de proporcionar información no consiste en revelar todo el algoritmo y menos aún su código fuente, o el secreto empresarial que lo protege. A la misma conclusión llega, por ejemplo, el MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL, *Información algorítmica en el ámbito laboral. cit.*, pp. 12-15, según el cual, «la obligación de información derivada de los artículos 13.2.f) y 14.2.g) 15.1.h) RGPD y 64.4.d) ET no puede interpretarse como la obligación empresarial de facilitar el código fuente del algoritmo». Comparten esta postura, RIVAS VALLEJO, P., «Herramientas desde el derecho antidiscriminatorio...», *cit.*, p. 552, NUÑEZ SEOANE, J., *op.cit.*, p. 307, o LAZCOZ MORATINOS, G., *op.cit.*, entre otros. Para la mayoría de la doctrina, este derecho de información no implica el derecho a acceder al algoritmo, no solo porque puede estar protegido por secreto empresarial, sino porque implicaría acceder a centenares de páginas de códigos indecifrables. Por el contrario, como precisa GINÉS i FABRELLAS, A., *op.cit.*, p. 319, implica el derecho a obtener información sobre las variables o métricas utilizadas por el algoritmo para la elaboración de perfiles o toma de decisiones automatizada, su ponderación en la ecuación final y las consecuencias que pueden derivarse para la persona. Como explica PALMA ORTIGOSA, A., «Decisiones automatizadas en el RGPD...», *cit.*, p.14/25, el responsable del tratamiento debe explicar, en la medida de lo posible, el funcionamiento del programa informático, aportando información útil para el interesado –como por ejemplo los factores que más han influido en la elaboración del perfil– que le permita comprender el proceso y pueda así ejercer el resto de los derechos que le reconoce el RGPD en su artículo 22.3. En la *Directiva (UE) 2019/2161 del Parlamento Europeo y del Consejo de 27 de noviembre de 2019 por la que se modifica la Directiva 93/13/CEE del Consejo y las Directivas 98/6/CE, 2005/29/CE y 2011/83/UE del Parlamento Europeo y del Consejo*, sobre la mejora de la aplicación y modernización de las normas de protección de los consumidores de la Unión, también se aclara que no debe exigirse a los comerciantes que revelen el funcionamiento detallado de sus mecanismos de clasificación (incluidos los algoritmos), sino sólo una descripción general de los principales parámetros que determinan la clasificación, si bien no es necesario que dicha descripción se presente de un modo personalizado para cada una de las consultas concretas efectuadas (vid., Considerandos 22 y 23).

⁴⁸⁵ La AEPD señala que según la RAE la palabra «*significativa*» denota «*Que da a entender o conocer con precisión algo*» y esto habría que interpretarlo como información que, proporcionada al interesado, le hace consciente del tipo de tratamiento que se está llevando a cabo con sus datos y que le proporciona certeza y confianza sobre sus resultados, lo que está relacionado con el concepto de «explicabilidad» del tratamiento mediante IA.

trinal, que dicha información no sería suficiente ni significativa si impidiera al interesado ejercitar los derechos y garantías que el propio art. 22 RGPD le otorga, esto es, oposición, revisión, alegación e impugnación de la decisión⁴⁸⁶.

Sobre el contenido concreto de este derecho es interesante la opinión del Abogado General Priit Pikamäe en el asunto C-634/21 y en los asuntos acumulados C-26/22 y C-64/22 SCHUFA Holding y otros, a la que el TJUE ha dado especial relevancia. Considera que *no es suficiente con dar una explicación genérica sobre cómo funciona el proceso de toma de decisión, sin indicar los datos específicos que se tienen de la persona y que han sido tenidos en cuenta para elaborar el informe negativo de solvencia*. A su juicio, «la obligación de proporcionar “información significativa sobre la lógica aplicada” debe entenderse en el sentido de que incluye *explicaciones suficientemente detalladas sobre el método utilizado para el cálculo del score y sobre las razones que han conducido a un resultado determinado*. En general, el responsable del tratamiento debería proporcionar al interesado información general, en particular sobre los *factores que han sido tenidos en cuenta en el proceso de toma de decisiones y sobre su importancia relativa desde el punto de vista agregado, que también le resulte útil para impugnar cualquier “decisión” en el sentido de la disposición del RGPD que consagra el “derecho” a no ser objeto de una decisión basada únicamente en un tratamiento automatizado, incluida la elaboración de perfiles*»⁴⁸⁷.

Dando un paso más, y sin perder de vista las premisas y consideraciones anteriores, la siguiente pregunta a resolver sería la siguiente: ¿sobre qué extremos concretos se debería informar al interesado? La AEPD manifiesta que, aunque la respuesta dependerá del tipo de componente de IA utilizado, un ejemplo de información que podría tener relevancia para el interesado, podría ser la siguiente: *«El detalle de los datos empleados para la toma de decisión, más allá de la categoría, y en particular información sobre los plazos de uso de los datos (su antigüedad); la importancia relativa que cada uno de ellos tiene en la toma de decisión; la calidad de los datos de entrenamiento y el tipo de patrones utilizados; los perfiles realizados y sus implicaciones; los valores de precisión o error según la métrica adecuada para medir la bondad de la inferencia; la existencia o no de supervisión humana cualificada; la referencia a auditorías, especialmente sobre las posibles desviaciones de los resultados de las inferencias, así como la certificación o certificaciones realizadas sobre el sistema de IA. En el caso de sistemas adaptativos o evolutivos, la última auditoría realizada; en el caso de que el sistema IA contenga información de terceros identifica-*

⁴⁸⁶ En este sentido, por ejemplo, NUÑEZ SEOANE, J., *op.cit.*, p. 307 y NAVAS NAVARRO, S., «La perspectiva de género...», *cit.*, p. 15/23.

⁴⁸⁷ Comunicado de prensa del TJUE n. 49/2023, 16 de marzo de 2023, con las conclusiones del Abogado General disponible en: https://curia.europa.eu/jcms/jcms/Jo2_7052/es/?annee=2023. Las cursivas del texto son nuestras

bles, la prohibición de tratar esa información sin legitimación y de las consecuencias de realizarlo»⁴⁸⁸.

Por su parte, el Ministerio de Trabajo y Seguridad Social de España publicó en 2022 una *Guía práctica sobre la obligación de los empresarios de informar a los trabajadores sobre el uso de algoritmos*⁴⁸⁹ que, aunque dirigida al ámbito de las relaciones laborales, creemos extrapolable a cualquier otro, al menos en lo relativo al contenido concreto de la información que deberá proporcionarse al interesado respecto del sistema algorítmico utilizado para la toma de decisiones que le afectan. En dicho documento se establece, con gran detalle, que deberá informarse de los siguientes extremos:

- a) *Información sobre el uso de algoritmos o sistemas de IA* para tomar decisiones automatizadas, incluyendo la elaboración de perfiles, identificando la tecnología utilizada por parte del algoritmo y las decisiones respecto de las que se utiliza tal tecnología. Esto incluiría, por ejemplo, informar sobre si la tecnología empleada genera un algoritmo de «caja negra» o si se trata de un algoritmo de aprendizaje continuo, el concreto *software* o producto utilizado y si cuenta, en su caso, con algún tipo de certificación, la empresa suministradora, si la empresa ha realizado alguna alteración sobre el producto, o el grado de intervención humana cualificada en las decisiones adoptadas, esto es, si las mismas se adoptarán mediante sistemas de IA o simplemente con apoyo en los resultados de dichos sistemas.
- b) *Información significativa, clara y simple sobre la lógica y funcionamiento del algoritmo*, que incluiría tanto *las variables tomadas en consideración* (entendidas como la información o factores utilizados por el algoritmo para tomar la decisión o la elaboración del perfil⁴⁹⁰, incluyendo si algunas de estas variables son datos personales), como *los parámetros utilizados* (entendidos como la ponderación relativa de cada variable en el modelo para la toma de la decisión, así como cualquier cambio de estos parámetros que modifique el comportamiento del algoritmo)⁴⁹¹. También habría que informar sobre las *reglas e instrucciones*

⁴⁸⁸ P. 24/53.

⁴⁸⁹ Dicha Guía, que fue elaborada por una comisión formada por expertos en algoritmos, concreta y sistematiza las obligaciones empresariales de información con el fin de garantizar, entre otros, el derecho fundamental a la no discriminación.

⁴⁹⁰ En caso de elaboración de perfiles, también hay que informar de la tipología de perfiles que elabora el algoritmo y, específicamente a la persona trabajadora ex artículos 13.2.f) y 14.2.g) RGPD, habría que darle *información sobre la asignación concreta a uno de ellos*.

⁴⁹¹ Por ejemplo, en un sistema algorítmico utilizado en un proceso de selección, las variables podrían ser la formación de la persona o la experiencia profesional previa, y los parámetros serían el peso relativo que tiene cada variable en la decisión final de selección o contratación (por ejemplo, el

utilizadas por el algoritmo, entendidas como las reglas programación (ya sea las programadas de forma expresa o las derivadas del aprendizaje del algoritmo) que conducen a la toma de la decisión⁴⁹² y sobre los *datos de entrenamiento* y, en su caso, *validación* (ya que estos también influyen en su lógica o instrucciones del algoritmo), su calidad (adecuación, pertinencia, minimización atendiendo a la finalidad para los que fueron obtenidos, etc.) y el *tipo de patrones identificados en los datos de entrenamiento*⁴⁹³. En casos de elaboración de perfiles, las métricas de precisión o error en las tareas automatizadas (clasificación, puntuación, regresión, ordenación...) de las personas en los distintos perfiles. Y, por último, la referencia a las *auditorías* realizadas (en el caso de sistemas adaptativos o evolutivos la última auditoría realizada) *especialmente sobre las posibles desviaciones de los resultados de las inferencias*, o la *evaluación de impacto* realizada por parte de la empresa o una tercera empresa respecto del algoritmo o sistema de decisión automatizada utilizado.

- c) Información sobre las *consecuencias que puede tener sobre la persona la decisión* adoptada (por ejemplo, obtener un puesto de empleo o determinar las concretas condiciones laborales)⁴⁹⁴.

modelo podría ordenar a las personas en atención a su experiencia profesional previa –valorada en un 40%–, formación –valorada en un 20%– etc., hasta alcanzar el 100%).

⁴⁹² En un sistema IA utilizado por una plataforma de reparto de comida a domicilio, las reglas e instrucciones podrían ser que la asignación de pedidos entre sus repartidores se realizara de la siguiente manera: (i) excluyendo a las personas que se encuentran a una distancia superior de 1 km del restaurante, (ii) puntuando en una escala de X a Y en un 50% a las personas que se encuentran a una distancia mínima de 1 km del restaurante y en otro 50% a las personas con una disponibilidad mínima de 10 horas en alta demanda en las últimas dos semanas y (iii) asignando el pedido a la persona que obtuviera una puntuación superior; así como las demás reglas de programación, tales como aplicar las anteriores instrucciones a las personas conectadas en ese momento, excluir a las personas que están en un pedido en curso, etc.

⁴⁹³ Por ejemplo, en un algoritmo para un proceso de selección la empresa tendría que comunicar que ha utilizado (i) los datos de la plantilla de la empresa de los últimos 10 años; (ii) obtenidos cumpliendo con las exigencias de protección de datos, específicamente la obligación de haber informado a estas personas que sus datos serían reutilizados para el entrenamiento de un algoritmo para elaborar perfiles; y (iii) informar que, analizando los datos de entrenamiento, ha identificado el patrón estadístico que, por ejemplo, las personas con una formación en A cumplen un X% mejor sus funciones para un puesto determinado, en comparación con aquellas personas que tienen una formación en B.

⁴⁹⁴ Además, en el ámbito laboral, debe informarse a los representantes de los trabajadores sobre el impacto que las decisiones adoptadas mediante algoritmos o sistemas de decisión automatizada tienen en materia de igualdad y no discriminación entre mujeres y hombres. Así se deriva del artículo 64.3 ET que reconoce el derecho de la representación legal de la plantilla de acceder a información referente a la aplicación en la empresa del derecho a la igualdad y no discriminación entre

En los casos en que las decisiones automatizadas exigieran la previa realización de una evaluación de impacto (art. 35 RGPD), el contenido de la información suministrada podría ser aún más exhaustivo, ya que sería ésta la que determinaría las «medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado»⁴⁹⁵.

Toda la información suministrada debería de cumplir, además, las exigencias formales y el nivel de transparencia del art. 12 RGPD, esto es, habría de facilitarse de forma clara, simple y comprensible a fin de que el interesado pudiera ejercer sus derechos. Como señala la *Carta ética europea sobre el uso de la IA en los sistemas judiciales*, «cuando se utiliza la inteligencia artificial, los derechos de los interesados son de especial importancia, y el control que cada uno de nosotros debe tener sobre nuestra información personal implica que debe ser posible ejercer los siguientes derechos: el derecho de los interesados a no estar sujeto a decisiones automatizadas que los afecten significativamente sin que se tenga en cuenta su punto de vista, el derecho a obtener información sobre el razonamiento subyacente al procesamiento de datos realizado por algoritmos, el derecho a oponerse a dicho procesamiento y el derecho a un recurso legal»⁴⁹⁶. Todo esto enfatiza la necesidad de transparencia en el tratamiento, en tanto en cuanto el interesado solo podrá impugnar la decisión o expresar su punto de vista si comprende plenamente cómo se ha tomado y sobre qué base.

Para garantizar la efectividad de dichos derechos, el RGPD confiere poderes de investigación y sanción a la autoridad de control (vid., arts. 53 y 83). Para un sector de la doctrina, la correlación de derechos individuales y de poderes o competencias de la AEPD es lo que realmente garantiza el cumplimiento de los deberes de *transparencia hacia el interesado*, y por ende, la efectivi-

mujeres y hombres. Así como del artículo 7 Real Decreto 901/2020, referente al diagnóstico en un contexto de elaboración de un Plan de Igualdad.

Vid., AEPD., *Adecuación...*, cit., pp. 24 y ss., en relación con el alcance de la información sobre la «importancia» y «consecuencias previstas» apunta que la misma debe versar «sobre el tratamiento previsto o futuro, y sobre cómo puede afectar la decisión automatizada al interesado».

⁴⁹⁵ Para NUÑEZ SEOANE, J., *op.cit.*, pp. 312-313, la evaluación de impacto es un recurso alternativo para que el responsable del tratamiento que por legítimos motivos no quiera o no pueda revelar el funcionamiento del algoritmo, pueda justificar y demostrar ante el interesado con plenas garantías que las decisiones automatizadas que adopte no supondrán un riesgo para sus derechos y libertades. Para RIVAS VALLEJO, P., «Herramientas desde el derecho antidiscriminatorio...», cit., p. 553, a diferencia de las auditorías del algoritmo, la evaluación de impacto permite actuar igualmente con carácter previo a la implantación de decisiones normativas o privadas que comporten el uso de algoritmos de decisión con el fin de evitar su impacto futuro y constituiría una barrera adicional a la auditoría de la IA y a la participación de la representación legal de los trabajadores ex art. 65.5 ET.

⁴⁹⁶ Pp. 40-41.

dad del derecho a una explicación de las personas afectadas por las decisiones automatizadas⁴⁹⁷.

En base a todos estos mimbres, se puede concluir que el RGPD se erige en uno de los principales instrumentos para hacer frente a la opacidad de los sistemas de toma de decisiones con impacto en las personas. Por un lado, porque trata de impedir el *primer nivel de opacidad* al que nos referimos anteriormente, cuando establece en los arts.13.2 f), 14.2 g) y 15.1 h) el derecho a ser informado de la *existencia de decisiones automatizadas*⁴⁹⁸. Así, los interesados no deberían ignorar, en teoría, que el proceso decisorio que les afecta se va a realizar por medio de un tratamiento automatizado de sus datos personales. El problema reside en que, en la práctica, no llegan a ser conscientes de ello porque los responsables del tratamiento no les informan de este extremo al considerar que su decisión no es automatizada, aunque de hecho sí lo sea⁴⁹⁹. Por otro lado, porque pretende evitar el *segundo nivel de opacidad*, ya que en tales casos deberá proporcionarse al interesado información significativa con el contenido y alcance anteriormente comentados. Por ende, puede afirmarse que en cierta medida dicha normativa contribuye a que el interesado pueda desafiar la validez de un algoritmo y pueda acceder a la ponderación que se le otorga a sus diversos elementos y a cualquier conclusión errónea a la que llegue.

Sin perjuicio de esto, la efectividad de este derecho de información reforzado reconocido en el RGPD y su correlativo deber de transparencia se ha puesto en duda por la doctrina cuando se emplean sistemas automatizados de decisiones basados en técnicas de IA como el *machine o deep learning*. Ello por diversas razones: por un lado, porque cuando se trata de algoritmos de *aprendizaje automático* que van modificándose progresivamente y de forma autónoma, es difícil proporcionar la información sobre la lógica del algoritmo por el efecto de caja negra que los caracteriza; por otro lado, porque como se explicó en su momento, en muchos casos se arguyen derechos de propiedad intelectual o secreto empresarial para no ofrecer información algorítmica

⁴⁹⁷ En este sentido, SÁEZ LARA, C., *op.cit.*, pp. 52 y ss.

⁴⁹⁸ LAZCOZ MORATINOS, G., *op.cit.*, p. 293.

⁴⁹⁹ Advierte de este hándicap PALMA ORTIGOSA, A., «Decisiones automatizadas en el RGPD...», *cit.*, p. 4/25, para quien, la indeterminación de los conceptos jurídicos recogidos en el art. 22 RGPD, conlleva que en muchas ocasiones los responsables desconocerán de buena voluntad si se trata de un tratamiento al que deban aplicarse las reglas del precepto. Además, teniendo en cuenta que este precepto establece mayores exigencias para ellos y más derechos en favor de los titulares de los datos, ante situaciones dudosas tenderán a huir de su aplicación por entender que el tratamiento que están llevando a cabo no se asimila al definido en el precepto. Esclarecer todos estos conceptos jurídicos ayudaría al conjunto de operadores jurídicos a determinar cuándo el tratamiento concreto que se está llevando a cabo se subsume en el art.22 y cuándo no.

a los interesados, no habiendo una postura doctrinal unánime acerca de la prevalencia o no de dichos derechos sobre los derechos o intereses de los individuos afectados por el tratamiento de sus datos personales cuando ambos entran en conflicto⁵⁰⁰. Por estas y otras razones un sector doctrinal considera que, si bien el art. 22 RGPD es útil en el contexto de la protección de datos, no lo es tanto para el control del uso de la IA aplicada a datos masivos. Así lo entiende REBOLLO DELGADO, para quien «dicho precepto aporta una solución tangencial a un problema sobrevenido, como es la aplicación de la IA a bases masivas de datos, por lo que su eficacia es limitada. Parece necesaria una regulación más específica, que concrete todas las posibilidades y lo haga desde la perspectiva de la protección del usuario, de la persona implicada. Pero aquí nos volvemos a encontrar con el problema recurrente de qué tipo de información ofrecemos, qué nivel técnico tiene, y en qué grado es entendible para el usuario o titular de los datos. Este problema de aplicabilidad no es menor, y plantea la vuelta a la dinámica de la Directiva 95/46/UE y que desterró el RGPD, es decir, que al responsable o encargado del tratamiento no le es suficiente con cumplir la norma, sino que debe conseguir la finalidad propuesta por ella. Esto supone un endoso al responsable o encargado del tratamiento de la teleología del mandato jurídico de la norma, cuya efectividad suele quedar difuminada»⁵⁰¹.

A nuestro modo de ver, en la medida en que hay consenso en que sería suficiente con compartir una explicación funcional en lenguaje común para que cualquier persona interesada, sin los conocimientos técnicos suficientes, pudiera entender bajo qué parámetros ha podido ser afectado por un modelo matemático de decisión, sin necesidad de revelar el código fuente o el secreto empresarial que lo protege⁵⁰², se evitarían en gran medida perjuicios a los derechos de propiedad intelectual o secretos empresariales y la negativa a proporcionar información carecería de justificación. Más aún si se tiene en cuenta que los expertos mantienen que existen métodos en la ciencia de la computación que permiten supervisar el modelo para averiguar si es necesario hacer ajustes o correcciones que no comprometen el contenido del algoritmo cuando está sujeto a derechos de propiedad intelectual o incluso por razones de seguridad⁵⁰³.

⁵⁰⁰ Sobre esta polémica puede verse LÓPEZ-TARRUELLA MARTÍNEZ, A., *op.cit.*, pp. 197-199.

⁵⁰¹ REBOLLO DELGADO, L., *op.cit.*, pp. 108.

⁵⁰² Tal como sostiene el GT29 en sus *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679*, p. 28. Y todo ello en la medida en que permita al interesado ejercitar los derechos de oposición, revisión, alegación e impugnación. Recuperado de <https://www.aepd.es/sites/default/files/2019-12/wp251rev01-es.pdf>

⁵⁰³ RIVAS VALLEJO, P., «Herramientas desde el derecho antidiscriminatorio...», *cit.*, p. 555, citando estudios sobre el tema.

Por último, no está claro cuáles serían las consecuencias de incumplir los deberes y garantías asociados a la automatización de decisiones, esto es, qué pasaría por ejemplo en caso de que el responsable no garantizara el derecho a obtener intervención humana o a obtener una explicación de la decisión tomada después de la evaluación o a conocer la lógica aplicada por el sistema. Más allá de la posible indemnización al interesado por los daños derivados de la vulneración de sus derechos en virtud del art. 82 RGPD y, en su caso, de las oportunas sanciones al responsable, habría que preguntarse si podría exigirse también la nulidad de la decisión automatizada. La duda se suscita porque el RGPD no prevé la nulidad como sanción ante la vulneración del derecho a la protección de datos, a diferencia de lo que ocurre, por ejemplo, en los casos de vulneración del derecho a la igualdad y no discriminación. A nuestro modo de ver, creemos que si la decisión automatizada fuera discriminatoria, sería posible el ejercicio de esta acción *ex* art. 26 de la Ley 15/2022; pero en caso contrario, y a la vista de que el RGPD no la establece, no creemos que pudiera solicitarse. Ello por cuanto, si bien no es necesario que una norma establezca expresamente la sanción de nulidad para que ésta se imponga, es posible que la propia norma recoja un efecto distinto para el caso de su contravención. Así ocurre, como se ha dicho, con la normativa de protección de datos.

No obstante, cuando fuera un acto de la Administración el que lesionase los derechos y libertades susceptibles de amparo constitucional, cabe recordar que el art. 47. 1. a) LPAC establece su nulidad de pleno derecho como sanción.

2.5.3. Otras normativas adoptadas a nivel europeo previas al Reglamento (UE) 2024/1689 (RIA)

Más allá del RGPD, a nivel europeo contamos con algunas normas que regulan y exigen transparencia algorítmica en ámbitos concretos. Respecto a las plataformas que prestan servicios de intermediación en línea, el *Reglamento (UE) 2019/1150 del Parlamento Europeo y del Consejo, de 20 de junio de 2019, sobre el fomento de la equidad y la transparencia para los usuarios profesionales de servicios de intermediación en línea* (conocido como Reglamento P2B), obliga a plataformas y motores de búsqueda a dar publicidad a los parámetros principales de funcionamiento de sus algoritmos de prelación de ofertas (arts. 5 y 7). El objetivo es que los usuarios (profesionales) de estas plataformas comprendan cómo funcionan los algoritmos que regulan los mecanismos de clasificación y puedan comparar las prácticas de clasificación empleadas por varios provee-

dores⁵⁰⁴. En esta línea, se afirma en el Considerando 24 que estos proveedores de servicios de intermediación en línea *«deben definir de antemano los parámetros principales que determinan la clasificación a fin de incrementar la previsibilidad para los usuarios profesionales, facilitarles un mayor entendimiento de cómo funciona el mecanismo de clasificación, y permitirles comparar las prácticas de clasificación de varios proveedores. La configuración específica de esta obligación de transparencia es importante para los usuarios profesionales, ya que implica el conocimiento de un conjunto limitado de parámetros que son los más pertinentes entre un número posiblemente mucho mayor de parámetros que influyen de alguna forma en la clasificación. [...] Debe entenderse que el concepto de parámetro principal hace referencia a los criterios generales, procesos, señales específicas incorporadas en los algoritmos u otros mecanismos de ajuste o degradación que se utilicen en la clasificación»*. No obstante, similarmente a lo que estableció el GT 29 respecto al RGPD, en el Considerando 27 se advierte que lo anterior no significa que estén obligados a revelar el funcionamiento detallado de sus mecanismos de clasificación, incluidos los algoritmos».

Por otro lado, la *Directiva (UE) 2019/2161 del Parlamento Europeo y del Consejo de 27 de noviembre de 2019 por la que se modifica la Directiva 93/13/CEE del Consejo y las Directivas 98/6/CE, 2005/29/CE y 2011/83/UE del Parlamento Europeo y del Consejo, en lo que atañe a la mejora de la aplicación y modernización de las normas de protección de los consumidores de la Unión*, establece que los comerciantes pueden personalizar el precio de sus ofertas para determinados consumidores o determinadas categorías de consumidores basándose en la toma de decisiones automatizada y la elaboración de perfiles a partir de su comportamiento, pero en tales casos se les deberá informar acerca de los principales parámetros por defecto que determinan la clasificación de las ofertas mostradas y su importancia relativa frente a otros parámetros, a fin de que puedan tener en cuenta los riesgos potenciales de su decisión de compra. Se aclara que este requisito de información se entiende sin perjuicio de la Directiva (UE) 2016/943 y que no debe exigirse a los comerciantes que revelen el funcionamiento detallado de sus mecanismos de clasificación (incluidos los algoritmos), sino sólo una descripción general de los principales parámetros que determinan la clasificación, si bien no es necesario que dicha descripción se presente de un modo personalizado para cada una de las consultas concretas efectuadas⁵⁰⁵.

⁵⁰⁴ Sobre este Reglamento y la discriminación algorítmica, puede verse ARGELICH COMELLES, C., «Deberes de transparencia del Reglamento 2019/1150 (P2B Regulation) para prevenir la discriminación algorítmica del consumidor en los sistemas de prelación de ofertas», *Cuadernos de derecho transnacional*, vol. 15, n. 1, 2023, pp. 129-135.

⁵⁰⁵ Así, vid., el Considerando 22 donde también se añade que «por parámetros que determinan la clasificación se entienden los criterios generales, procesos, señales específicas incorporadas en

También en muchos instrumentos de *soft law* provenientes de la UE se aboga por la transparencia, información y explicabilidad de los sistemas algorítmicos. Por ejemplo, y sin perjuicio de los ya mencionados en el apartado relativo a la opacidad del código fuente, en la *Resolución de 12 de febrero de 2020 sobre los procesos automatizados de toma de decisiones*, el Parlamento Europeo considera que cuando un consumidor esté interactuando con un sistema que automatice la toma de decisiones, se le debe informar tanto de su funcionamiento (segundo nivel de opacidad), como de la manera de contactar con un ser humano con poder de decisión y de cómo se pueden verificar y corregir las decisiones del sistema. Estos deberes implican lógicamente, como paso previo, haber informado al consumidor del uso del sistema algorítmico para decidir sobre el asunto concreto (primer nivel de opacidad). También en su *Resolución sobre aspectos éticos de la IA de 20 de octubre de 2020*, estima que los consumidores deben tener derecho a ser debidamente informados de manera comprensible, oportuna, normalizada, exacta y accesible de la existencia, el razonamiento, el posible resultado y las repercusiones que puedan tener para ellos los sistemas algorítmicos, así como de la manera de contactar con un ser humano con poder de decisión y de cómo se pueden verificar, impugnar de manera útil y corregir las decisiones del sistema. E incluso en la *Propuesta de Reglamento que acompañaba a la Resolución de 20 octubre de 2020 en materia de responsabilidad civil* se establecía que, tanto el operador al que se considerara responsable de los daños ocasionados por el sistema de IA como la persona afectada, podrían usar los datos generados por el mismo, bien para que el primero pudiera demostrar la culpa concurrente de la persona afectada por el sistema, bien para que la segunda pudiera probar o aclarar los presupuestos de su demanda de responsabilidad civil, siempre y cuando este acceso a los datos se hiciera de conformidad con el Reglamento (UE) 2016/679 y otras leyes en materia de protección de datos⁵⁰⁶.

En definitiva, las instituciones europeas han venido manteniendo que los consumidores tienen derecho a saber *cuándo y cómo* se utilizan los sistemas de IA en procesos que les afectan y que las empresas deben proporcionar *información clara sobre cómo se toman las decisiones automatizadas* y cómo se recopilan y utilizan los datos de los consumidores en los que basan sus decisiones.

Por último, contamos en el ámbito europeo con un reconocimiento de derechos digitales del que se desprende el derecho a conocer que estamos siendo objeto decisiones algorítmicas. Se trata de la *Declaración sobre los Derechos y Principios Digitales* adoptada bajo la forma de declaración conjunta y solemne por el Parlamento Europeo, el Consejo y la Comisión el 23 de enero de 2023,

los algoritmos u otros mecanismos de ajuste o degradación que se utilicen en la clasificación», y el Considerando 23.

⁵⁰⁶ Así lo recogía su art. 10.2.

donde, respecto a las interacciones con algoritmos y sistemas de IA, se recoge el compromiso de dichos órganos de «velar por un nivel adecuado de *transparencia en el uso de los algoritmos y la inteligencia artificial* y porque las personas estén *informadas y capacitadas para utilizarlos cuando interactúen con ellos*». Aunque se trata de un texto puramente declarativo que no afecta a las normas de la UE ni a su aplicación, las instituciones europeas pretenden ensalzar y recordar principios y valores que informen y orienten las decisiones de los agentes digitales y de los responsables políticos sobre el uso de la IA.

2.5.4. El Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial

La transparencia ocupa un lugar destacado en el RIA y se refiere tanto al grado de implantación en un sistema (finalidad, diseño, función y datos) como al modo en que se utiliza y a las repercusiones que genera⁵⁰⁷. Además, aunque el RIA aboga principalmente por la *transparencia interna* (sobre todo hacia los responsables del despliegue), también establece algunas obligaciones a favor de la *transparencia externa* y ha acabado incorporando, como novedad destacable respecto a su versión inicial, el derecho a una explicación de las personas afectadas por las decisiones algorítmicas.

2.5.4.1. Obligaciones de transparencia y comunicación de información

A) Sistemas de alto riesgo

Cuando los sistemas no son lo suficientemente transparentes y explicables, ni están bien documentados, pueden impedir el ejercicio de derechos procesales fundamentales, como el derecho a la tutela judicial efectiva y a un juez imparcial, el derecho de defensa y la presunción de inocencia⁵⁰⁸. Para evitar la conculcación de tales derechos, el RIA establece, en su art. 13.1, que los sistemas de alto riesgo «se diseñarán y desarrollarán de un modo que garantice que funcionan con un nivel de transparencia suficiente para que los responsables del despliegue interpreten y usen correctamente su información de salida. Se garantizará un tipo y un nivel de transparencia adecuados para que el proveedor y el responsable del despliegue cumplan las obligaciones oportunas previstas en la sección 3». De esta disposición pueden extraerse varias ideas:

⁵⁰⁷ FERNÁNDEZ ROZAS, J.C., *op.cit.*, p. 25.

⁵⁰⁸ Así se advierte en el Considerando 59 RIA.

En primer lugar, en cuanto al *grado de transparencia* requerido para subsanar la opacidad de algunos sistemas, se exige una transparencia *suficiente* y no total⁵⁰⁹, con el fin de evitar colisiones con los derechos de propiedad intelectual. Con esta declaración, la Comisión se posiciona claramente a favor del equilibrio entre la necesidad de transparencia y los intereses empresariales, en la misma línea que ya venía manteniendo el Consejo de Europa. La versión inicial del RIA indicaba expresamente en su Exposición de Motivos que *ni siquiera las obligaciones que exigían una mayor transparencia debían afectar de manera desproporcionada a los derechos de propiedad intelectual* puesto que únicamente debían aplicarse a la información mínima necesaria para que las personas pudieran ejercer su derecho a una compensación efectiva y solo exigían la transparencia necesaria hacia las autoridades de supervisión y las encargadas de la aplicación de la ley. También se establecía que la información habría de divulgarse siempre con arreglo a la legislación pertinente en la materia, y que, tanto las autoridades públicas como los organismos notificados tendrían que cumplir obligaciones de confidencialidad vinculantes en los casos en que debieran tener acceso a información confidencial o al código fuente para examinar el cumplimiento de las obligaciones sustanciales⁵¹⁰. En la versión definitiva del RIA también se establece reiteradamente que toda la información o documentación a la que tengan acceso las autoridades, organismos notificados y cualquier otra persona física o jurídica que participe en su aplicación, deberá ser tratada de conformidad con las obligaciones de confidencialidad establecidas en el artículo 78.

En segundo lugar, el *tipo y nivel de transparencia* exigible ha de ser compatible con el cumplimiento de las obligaciones legales de proveedores y responsables del despliegue.

⁵⁰⁹ Para HUERGO LORA, A. «El proyecto de Reglamento sobre la Inteligencia Artificial», 2021, disponible en <https://almacenederecho.org/el-proyecto-de-reglamento-sobre-la-inteligencia-artificial> (última consulta 17/7/2022), la regulación de la transparencia es una de las cuestiones más delicadas, y el proyecto busca un equilibrio en virtud del cual el proveedor debe mostrar cómo funciona la aplicación, incluida su «lógica general», así como los presupuestos de partida o una descripción de los datos que se hayan utilizado para su creación, pero no se le exige transparencia total sobre el *software* utilizado. Considera el autor, que «estos requisitos son, en cierto modo, “objetivos” máximos a los que se debe tender, pero que pueden conseguirse de muchas maneras y también con niveles de intensidad diferentes». En este sentido se pronuncia la *Recomendación sobre la Ética de la Inteligencia Artificial de la UNESCO* (2021), al considerar que el grado de transparencia y explicabilidad debería ser siempre adecuado al contexto y al efecto, ya que puede ser necesario encontrar un equilibrio entre la transparencia y la explicabilidad y otros principios como la privacidad, la seguridad y la protección (ap. 38).

⁵¹⁰ Apartado 3.5 de la Exposición Motivos que contenía la Propuesta de RIA de 21 de abril de 2021 –COM(2021) 206 final–.

En tercer lugar, el RIA *asocia la transparencia con la información que se debe suministrar al responsable del despliegue del sistema*. Es decir, en principio la obligación de transparencia y comunicación de información va dirigida a informar a los agentes que implementan la tecnología y no a las personas potencialmente afectadas por la decisión automatizada. Por tanto, se trata de una transparencia interna y no hacia los usuarios finales.

El Considerando 72 RIA da amplia cuenta de que el objetivo de la transparencia exigible a los sistemas de alto riesgo es que los responsables del despliegue puedan «*comprender la manera en que el sistema de IA funciona, evaluar su funcionalidad y comprender sus fortalezas y limitaciones*», así como «*utilizar el sistema y tomar decisiones con conocimiento de causa*». No obstante, como se explicará *infra*, esto repercutirá indirecta y favorablemente en la transparencia hacia los usuarios finales –transparencia externa– en la medida en que contribuirá a dar cumplimiento a las exigencias del art. 86 que obliga a los responsables a dar explicaciones a los afectados por la decisión del sistema.

Con estos fines, los sistemas de IA de alto riesgo deben ir acompañados de la documentación y las instrucciones de uso oportunas (en un formato digital o de otro tipo adecuado) e incluir información concisa, completa, correcta y clara que sea pertinente, accesible y comprensible para los responsables del despliegue, en particular sobre los posibles riesgos para los derechos fundamentales y de discriminación, cuando corresponda. Además, para facilitar que los responsables del despliegue comprendan dichas instrucciones, deberán contener ejemplos ilustrativos según proceda⁵¹¹.

Concretamente, en las instrucciones de uso el proveedor de la tecnología quedará obligado a informar al responsable del despliegue de forma adecuada y suficiente sobre una serie de extremos *mínimos* que se recogen en el art. 13.3, entre los que se incluyen: A) la identidad y los datos de contacto del proveedor y, en su caso, de su representante autorizado; B) las características, capacidades y limitaciones del funcionamiento del sistema de IA de alto riesgo, y en particular: i) su finalidad prevista, incluido el entorno geográfico, funcional o de comportamiento específico en el que se pretende utilizar el sistema de IA de alto riesgo; ii) el nivel de precisión (incluidos los parámetros para evaluarla), solidez y ciberseguridad mencionado en el artículo 15 con respecto al cual se haya probado y validado el sistema de IA de alto riesgo y que puede esperarse, así como las circunstancias conocidas y previsibles que podrían afectar al nivel de precisión, solidez y ciberseguridad esperado; iii) cualquier circunstancia conocida o previsible, asociada a la utilización del sistema de IA de alto riesgo conforme a su finalidad prevista o a un uso indebido

⁵¹¹ Considerando 72.

razonablemente previsible, que pueda dar lugar a riesgos para la salud y la seguridad o los derechos fundamentales a que se refiere el art. 9.2; iv) en su caso, las capacidades y características técnicas del sistema de IA de alto riesgo para proporcionar información pertinente para explicar su información de salida; v) cuando proceda, su funcionamiento con respecto a personas o grupos de personas específicos en relación con los que se pretenda utilizar el sistema; vi) cuando proceda, especificaciones relativas a los datos de entrada, o cualquier otra información pertinente en relación con los conjuntos de datos de entrenamiento, validación y prueba usados, teniendo en cuenta la finalidad prevista del sistema de IA; vii) en su caso, información que permita a los responsables del despliegue interpretar la información de salida del sistema de IA de alto riesgo y utilizarla adecuadamente; C) los cambios en el sistema de IA de alto riesgo y su funcionamiento predeterminados por el proveedor en el momento de efectuar la evaluación de la conformidad inicial, en su caso; D) las medidas de vigilancia humana a que se hace referencia en el artículo 14, incluidas las medidas técnicas establecidas para facilitar la interpretación de la información de salida de los sistemas de IA por parte de los responsables del despliegue; E) los recursos informáticos y de *hardware* necesarios, la vida útil prevista del sistema de IA de alto riesgo, así como las medidas de mantenimiento y cuidado necesarias (incluida su frecuencia) para garantizar el correcto funcionamiento de dicho sistema, también en lo que respecta a las actualizaciones del *software*; F) cuando proceda, una descripción de los mecanismos incluidos en el sistema de IA para permitir a los responsables del despliegue recabar, almacenar e interpretar correctamente los archivos de registro de conformidad con el art. 12.

Por otro lado, otro requisito exigible por el RIA que también contribuye a la transparencia externa, es que los sistemas de alto riesgo deban *registrarse en una base de datos de la UE* que la Comisión gestionará con el propósito de redoblar la transparencia y vigilancia públicas y de fortalecer la supervisión *ex post* por parte de las autoridades competentes (art. 49 en relación con el art. 71)⁵¹². Como dispone el art. 71.4, la información de esta base de datos registrada de conformidad con lo dispuesto en el artículo 49 será accesible y estará a disposición del público de manera sencilla.

Por último, respecto a los *modelos de IA de uso general*, hay que tener en cuenta que los proveedores e implantadores quedan sometidos a obligaciones de transparencia más exigentes en virtud del RIA.

⁵¹² Que el objetivo de esta obligación es aumentar la transparencia hacia el exterior o de cara al público lo expresa claramente el Considerando 131: «Con el objetivo de facilitar la labor de la Comisión y de los Estados miembros en el ámbito de la IA, así como de incrementar la transparencia de cara al público ...».

B) *Determinados sistemas de IA con independencia de su riesgo*

El art. 50 impone a proveedores y responsables del despliegue obligaciones de transparencia específicas respecto a ciertos sistemas de IA, con independencia de si son o no de alto riesgo⁵¹³. Se trata de aquellos destinados a interactuar directamente con personas físicas, sistemas *que generen contenido sintético de audio, imagen, vídeo o texto*, sistemas de reconocimiento de emociones o de categorización biométrica y sistemas que generen o manipulen imágenes o contenidos de audio o vídeo *que constituyan una ultrafalsificación, o generen o manipulen texto que se publique para informar sobre asuntos de interés público*⁵¹⁴. En todos estos casos, se exceptúan de esta obligación los sistemas de IA autorizados legalmente para detectar, prevenir, investigar o enjuiciar delitos.

El objetivo que se persigue en estos casos con la transparencia e información es evitar la confusión o engaño de quienes interactúan o se ven expuestos a dichos sistemas⁵¹⁵. Para ello, en el primer caso, se deberá comunicar a las personas que están interactuando con un sistema de IA y no con un humano (excepto cuando resulte evidente *desde el punto de vista de una persona física razonablemente informada, atenta y perspicaz, teniendo en cuenta las circunstancias y el contexto de utilización*). Al aplicar dicha obligación, deben tenerse en cuenta las características de las personas físicas pertenecientes a colectivos vulnerables debido a su edad o discapacidad, en la medida en que el sistema de IA esté destinado a interactuar también con dichos colectivos.

En el segundo caso, se velará por que la información de salida del sistema esté marcada en un formato legible por máquina y que sea posible detectar que ha sido generada o manipulada de manera artificial y no por un ser humano. Se exige a los proveedores que integren soluciones técnicas y métodos para lograr este cometido que deberán ser lo suficientemente fiables, interoperables, eficaces y sólidos, en la medida en que sea técnicamente viable, teniendo en cuenta las técnicas disponibles o una combinación de dichas téc-

⁵¹³ En los Considerandos 132, 133 y 134 se justifica la imposición de tales obligaciones a estos sistemas.

⁵¹⁴ Apunta PEGUERA POCH, M., «Perspectiva y categorización del riesgo en el Reglamento de Inteligencia Artificial», en J.A. Moreno Martínez/P.J. Femenía López (coords.), *Inteligencia artificial y derecho de daños: Cuestiones actuales: Acorde al Reglamento (UE) 2024/1689*, Madrid, 2024, p. 926, que en rigor dichos sistemas no son de «riesgo limitado» porque pueden ser, a su vez, de alto riesgo, en cuyo caso deberán cumplir los correspondientes requisitos de modo acumulativo.

⁵¹⁵ Señala FERNÁNDEZ ROZAS, J.C., *op.cit.*, p. 16, que con los requisitos de transparencia que se imponen a los sistemas que interactúan con seres humanos se pretende dar lugar a un cambio significativo en las prácticas de divulgación a escala global en sitios web y aplicaciones, y garantizará que los usuarios estén informados sobre el funcionamiento y el propósito de los sistemas de IA con los que interactúan.

nicas, como marcas de agua, identificación de metadatos, métodos criptográficos para demostrar la procedencia y la autenticidad del contenido, métodos de registro, impresiones dactilares u otras técnicas, según proceda. Además, a la hora de aplicar esta obligación, deberán tener en cuenta las particularidades y limitaciones de los diversos tipos de contenido, los costes de aplicación y el estado actual de la técnica generalmente reconocido, según se refleje en las normas técnicas pertinentes⁵¹⁶.

En el tercer caso, se notificará a las personas físicas que están expuestas a sistemas de IA que tratan sus datos biométricos y que pueden determinar o inferir sus emociones o intenciones, así como incluirlas en categorías específicas en función del sexo, edad, color del pelo o de ojos, tatuajes, rasgos personales, origen étnico o preferencias e intereses personales. Dichos datos personales se tratarán de conformidad con el Reglamento y la Directiva de protección de datos personales. Además, si las personas cuyos datos se tratan son personas con discapacidad, toda la información y notificaciones deberán facilitarse en formatos accesibles.

Por lo que se refiere al último tipo de sistemas mencionado, deberá hacerse público que los contenidos, imágenes o texto han sido generados o manipulados de manera artificial etiquetando los resultados de salida generados por la IA e indicando su origen artificial. No obstante, cuando el contenido forme parte de una obra o programa manifiestamente creativos, satíricos, artísticos, de ficción o análogos, sólo se deberá hacer pública la existencia de dicho contenido generado o manipulado artificialmente de una manera adecuada que no dificulte la exhibición o el disfrute de la obra, su explotación y uso normales, al tiempo que se conserva su utilidad y calidad⁵¹⁷.

⁵¹⁶ Indica el Considerando 133 que «Dichas técnicas y métodos pueden implantarse a nivel de sistema de IA o a nivel de modelo de IA, incluidos modelos de IA de uso general que generan contenidos, facilitando así el cumplimiento de esta obligación por parte del proveedor posterior del sistema de IA. Para garantizar la proporcionalidad, conviene prever que esta obligación de marcado no se aplique a los sistemas de IA que desempeñen una función de apoyo a la edición estándar o no alteren sustancialmente los datos de entrada facilitados por el responsable del despliegue o su semántica».

⁵¹⁷ Se precisa con más detalle en el Considerando 134, que «el cumplimiento de esta obligación de transparencia no debe interpretarse como un indicador de que la utilización del sistema de IA o de sus resultados de salida obstaculiza el derecho a la libertad de expresión y el derecho a la libertad de las artes y de las ciencias, garantizados por la Carta, en particular cuando el contenido forme parte de una obra o programa manifiestamente creativos, satíricos, artísticos, de ficción o análogos, con sujeción a unas garantías adecuadas para los derechos y libertades de terceros [...] Además, también conviene prever una obligación de divulgación similar en relación con el texto generado o manipulado por una IA en la medida en que se publique con el fin de informar al público sobre asuntos de interés público, a menos que el contenido generado por la IA haya sido sometido a un proceso de revisión humana o de control editorial y que una persona física o jurídica ejerza la responsabilidad editorial de la publicación del contenido».

Se exige inmediatez a la hora de revelar la información y que la misma se ajuste a los requisitos de accesibilidad aplicables. Así, toda información y notificaciones deberán facilitarse de manera clara y distinguible a más tardar con ocasión de la primera interacción o exposición al sistema (art. 50.5).

2.5.4.2. *Derecho a una explicación*

En su versión inicial de 21 de abril de 2021, el RIA no reconocía ningún derecho a ser informado de la existencia de decisiones automatizadas, ni un derecho de acceso a los datos o información del sistema, ni un derecho a la explicabilidad por su mal funcionamiento⁵¹⁸. Es decir, no se recogían en dicho instrumento derechos subjetivos para los afectados por dichas decisiones (usuarios finales), sino que sus previsiones iban dirigidas exclusivamente a los proveedores de tecnología basada en IA respecto a los responsables de su despliegue.

Esta carencia fue muy criticada por la doctrina que la calificó como un punto ciego de la regulación⁵¹⁹ y el Parlamento Europeo introdujo una enmienda al texto de la propuesta inicial en la que recogía un *derecho de*

⁵¹⁸ Así lo critica, NAVAS NAVARRO, S., *Daños ocasionados por sistemas de inteligencia artificial...*, cit., p. 73. Con las medidas recogidas tanto en la *Propuesta de Directiva para adaptar las normas de responsabilidad extracontractual en la Inteligencia Artificial (AI Liability Directive)*, como en la *Propuesta de Directiva para modernizar el régimen de responsabilidad por productos defectuosos*, ambas de 28 de septiembre de 2022, la Comisión Europea trató de suplir esa carencia inicial del RIA mediante el establecimiento de un derecho de acceso a la información y a los datos que obran en poder del demandado o potencial demandado acerca de ese sistema de IA (arts. 3 y 8 respectivamente). Por tanto, se eliminaban los obstáculos con los que podían encontrarse las víctimas para acceder a información necesaria del sistema que le permitiera probar los presupuestos para reclamar responsabilidad civil por los daños sufridos. Se reconoce un derecho subjetivo a la exhibición de pruebas similar al derecho de información y acceso que tiene el afectado por un tratamiento de datos en el RGPD, sometido no obstante a ciertas condiciones que después se explicarán. El principal objetivo de dicha medida era incentivar a proveedores y responsables del despliegue de los sistemas a facilitar información sobre los mismos a los perjudicados y a cumplir las medidas del RIA dirigidas a garantizar la transparencia.

⁵¹⁹ Así, ÁLVAREZ CUESTA, H. «Inteligencia artificial: derecho de la UE y derecho comparado. La propuesta de una ley sobre IA», cit. Otros autores que critican dicha omisión del RIA son, por ejemplo, RIVAS VALLEJO, P., «Herramientas desde el derecho antidiscriminatorio...», cit., p. 542; MUÑOZ GARCÍA, C., «Adaptar o reformular la directiva 85/374 sobre responsabilidad por daños causados por productos defectuosos a la inteligencia artificial», accesible en <https://diariolaleylaleynext.es/dll/2022/03/01/adaptar-o-reformular-la-directiva-85-374-sobre-responsabilidad-por-danos-causados-por-productos-defectuosos-a-la-inteligencia-artificial/>; y TAPIA HERMIDA, A., «Propuesta de Directiva sobre responsabilidad en Inteligencia Artificial Responsable (IAR)», en <https://sanchezcalero.com/propuesta-de-directiva-sobre-responsabilidad-en-inteligencia-artificial-responsable-iar/> (última consulta: 17 de enero de 2025), para quien la laguna regulatoria surge porque los requisitos establecidos en Ley de IA no ofrecen ayuda individual a quienes han sufrido daños causados por la IA.

explicabilidad algorítmica muy similar al consagrado en el RGPD, pero con la importante diferencia de ser aplicable también a las *decisiones semiautomatizadas*, esto es, a aquellas que adopta una persona con apoyo en los resultados del sistema. De hecho, en el art. 68 *quarter*, titulado «Derecho a explicación de la toma individual de decisiones», se establecía que «toda persona afectada sujeta a una *decisión adoptada por el implementador* [ahora responsable del despliegue] *sobre la base de la información de salida de un sistema de IA de alto riesgo* que produzca *efectos jurídicos o que le afectan significativamente* de una manera que considera que perjudica a su salud, seguridad, derechos fundamentales, bienestar socioeconómico o cualquier otro de sus derechos derivados de las obligaciones establecidas en el presente Reglamento, *tendrá derecho a solicitar al implementador una explicación clara y significativa, de conformidad con el artículo 13, apartado 1, sobre el papel del sistema de IA en el procedimiento de toma de decisiones, los principales parámetros de la decisión adoptada y los datos de entrada correspondientes*». Añadiendo en el apartado tercero que dicho precepto se aplicaría *sin perjuicio de lo previsto en los artículos 13, 14, 15 y 22 del RGPD*⁵²⁰.

El objetivo de dicha enmienda era subsanar la laguna de la que adolecía la normativa y conseguir que la transparencia algorítmica no sólo fuera dirigida a los operadores de la tecnología sino también a las personas físicas afectadas por las decisiones automatizadas. Tras dicha enmienda, la versión del texto aprobada el 13 de marzo de 2024 por la Eurocámara⁵²¹ recogió por primera vez ese *derecho de las personas afectadas a obtener una explicación de las decisiones individuales*, novedad importante respecto de la que no hubieron modificaciones tras su aprobación definitiva en mayo de ese mismo año.

Así, el actual art. 86 RIA dispone que «Toda persona que se vea afectada por una decisión que el responsable del despliegue adopte basándose en los resultados de un sistema de IA de alto riesgo que figure en el anexo III, con excepción de los sistemas enumerados en su punto 2, y que produzca efectos jurídicos o le afecte considerablemente del mismo modo, de manera que considere que tiene un efecto perjudicial para su salud, su seguridad o sus

⁵²⁰ El apartado segundo del art.68 *quarter* establecía excepciones a la aplicación de este derecho de explicación a disponer que «El apartado 1 no se aplicará a la utilización de sistemas de IA para los que el Derecho nacional o de la Unión prevea excepciones o restricciones a la obligación prevista en el apartado 1 en la medida en que dichas excepciones o restricciones respeten la esencia de los derechos y libertades fundamentales y sean una medida necesaria y proporcionada en una sociedad democrática»

⁵²¹ P9TA(2024)0138. Resolución legislativa del Parlamento Europeo, de 13 de marzo de 2024, sobre la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión (COM(2021)0206 - C9-0146/2021 - 2021/0106(COD)).

derechos fundamentales, tendrá derecho a obtener del responsable del despliegue explicaciones claras y significativas acerca del papel que el sistema de IA ha tenido en el proceso de toma de decisiones y los principales elementos de la decisión adoptada⁵²². Se consagra expresamente un derecho de explicación *ex post*, central en la ordenación de la IA, que viene a completar el limitado y criticable ámbito del art. 22 RGPD por cuanto se aplica a los sistemas de apoyo a la decisión y no sólo a los que tomen decisiones de forma totalmente automatizada.

Se establecen dos límites a la aplicación de este derecho: el primero, que no se aplicará cuando se utilicen sistemas de IA para los que existan excepciones o restricciones a la obligación prevista en el apartado 1 derivadas del Derecho nacional o de la UE (art. 86.2); el segundo, que solo se aplicará en la medida en que no esté ya previsto en el Derecho de la Unión (art. 86.3), disposición que contiene una clara referencia al RGPD.

Este derecho se completa con la obligación de los responsables del despliegue establecida en el nuevo art. 26, apartado 11, en virtud de la cual, en el caso de que se estén utilizando sistemas de IA de alto riesgo del anexo III que tomen decisiones o ayuden a tomar decisiones relacionadas con personas físicas, deberán informar a dichas personas de que están expuestas a la utilización de tales sistemas, sin perjuicio de lo dispuesto en el art. 50 sobre obligaciones de transparencia específicas cuando se usen determinados sistemas IA destinados a interactuar con personas físicas⁵²³. Por tanto, es evidente que los responsables del despliegue asumen un papel fundamental a la hora de garantizar la transparencia externa hacia las personas afectadas por los sistemas tanto antes (*ex ante*) como después de adoptada la decisión (*ex post*). Respecto a la primera, porque deben informar a las personas físicas de que son objeto de la utilización de un sistema de IA de alto riesgo, información que debe incluir tanto la finalidad prevista como el tipo de decisiones que se toman. Respecto a la segunda, porque deben informar de que gozan de un derecho a obtener una explicación de la decisión y darle cumplimiento.

De todo esto se infiere que, igual que dijimos respecto del RGPD, el RIA también pretende superar los dos niveles de opacidad algorítmica a los que se hizo referencia en el apartado 2. Por un lado, garantizando la transparencia

⁵²² El apartado segundo dispone que «No se aplicará el apartado 1 a la utilización de sistemas de IA para los que existan excepciones o restricciones a la obligación prevista en el apartado 1 derivadas del Derecho de la Unión o nacional de conformidad con el Derecho de la Unión». Y en el apartado tercero se establece que «El presente artículo se aplicará únicamente en la medida en que el derecho a que se refiere el apartado 1 no esté previsto de otro modo en el Derecho de la Unión».

⁵²³ También dispone el art. 26.11 que «En el caso de los sistemas de IA de alto riesgo que se utilicen a los efectos de la aplicación de la ley, se aplicará el artículo 13 de la Directiva (UE) 2016/680».

sobre el uso de dichos sistemas para que los ciudadanos sean conscientes de que es una IA la que decide sobre su persona; por otro lado, garantizando la transparencia sobre el contenido y justificación de la decisión final.

Por lo que se refiere al contenido concreto de la explicación que debe ofrecer el responsable al afectado por la decisión, dos son los extremos sobre los que ha de informarle: en primer lugar, del papel que el sistema de IA ha tenido en el proceso de toma de decisiones; y en segundo lugar, de los principales elementos de la decisión adoptada. Deberá hacerlo de forma clara y significativa, es decir, de forma suficiente para que el afectado pueda ejercer sus derechos.

Dar explicaciones sobre el *papel que ha jugado el sistema* a la hora de tomar la decisión implicará informar al interesado sobre si se trata de una decisión plena o semiautomatizada, es decir, si el resultado se ejecutó de forma totalmente automática y sin intervención humana, o si se utilizó como mero apoyo de la decisión adoptada finalmente por una persona. Además, en este último caso tendrá que indicar también si el resultado del sistema fue *determinante o no* para la toma de la decisión final y en qué medida lo fue⁵²⁴.

De hecho, la cuestión de hasta qué punto la salida o resultado del sistema de IA es *determinante* para la decisión finalmente adoptada por el usuario, ya no es solo relevante a los efectos de aplicar las garantías del art. 22 RGPD, sino que ahora también es decisiva para determinar si un sistema de IA es considerado de alto riesgo y, en consecuencia, para la aplicación de todas las garantías previstas en el RIA⁵²⁵. Es decir, los sistemas de IA independientes se clasifican como de alto riesgo si, a la luz de su *finalidad prevista*, presentan un *alto riesgo de ser perjudiciales para la salud y la seguridad o los derechos fundamentales de las personas*, teniendo en cuenta tanto la *gravedad* del posible perjuicio como la *probabilidad* de que se produzca, y se utilizan en una serie de ámbitos especificados en el Reglamento. No obstante, tal como se explica en el Considerando 53, pueden existir casos en los que, a pesar de utilizarse en esos ámbitos predefinidos, los sistemas no entrañen un riesgo considerable de causar perjuicios a intereses jurídicos protegidos *si no influyen sustancialmente en la toma de decisiones*. Se entenderá que no influyen sustancialmente en la toma de decisiones cuando no afecten al fondo, ni por consiguiente al resultado, de la toma de decisiones humana o automatizada. Las condiciones (una o varias) para que pueda considerarse que un sistema no plantea dicho riesgo al no influir sustancialmente en el resultado de la toma de decisiones, se recogen en el

⁵²⁴ Sobre el carácter «determinante» o no del resultado del sistema para la toma de la decisión final nos remitimos a las apreciaciones realizadas al comentar *supra* la STJUE de 7 de diciembre de 2023.

⁵²⁵ Así lo aprecia con acierto COTINO HUESO, L., «La primera sentencia del Tribunal de Justicia de la Unión Europea...», *cit.*, para quien, la relevancia del resultado del sistema de IA en la decisión finalmente adoptada ha pasado a ser en el RIA un elemento esencial para determinar si el sistema es de alto riesgo.

art. 6.3. Su concurrencia implicará que esos sistemas no queden sometidos a los requisitos que el RIA impone a los sistemas de alto riesgo.

Respecto a los *principales elementos de la decisión adoptada*, creemos que podría aplicarse aquí lo explicado en el apartado sobre el contenido del derecho de información reforzado en el RGPD⁵²⁶. Así, habría que informar, entre otras cosas, de los factores utilizados por el algoritmo para tomar la decisión o la elaboración del perfil, la ponderación relativa de cada variable en el modelo para la toma de la decisión y cualquier cambio de estos parámetros que modifique el comportamiento del algoritmo, las reglas e instrucciones utilizadas por el algoritmo, así como los datos utilizados por el sistema respetando en todo caso el RGPD.

A nuestro modo de ver, debe aplaudirse que el RIA haya acabado consagrando este derecho a favor de los potenciales afectados por las decisiones de la IA en la medida en que contribuirá a que comprendan cómo se adoptan y cómo pueden afectarles. A este fin contribuye también otra de las exigencias del RIA recogida en su art. 4 a la que haremos referencia al hablar del papel de la ética en la prevención de la discriminación, como es la de educar y formar en competencias de IA a todos los que operan con ella, con la que se pretende dotar a proveedores, responsables del despliegue y personas afectadas, de los conceptos necesarios para tomar decisiones con conocimiento de causa en relación con los sistemas de IA⁵²⁷.

2.5.5. Medidas adoptadas a nivel interno

2.5.5.1. Derecho privado español

Por lo que se refiere al Derecho privado español, y más concretamente al Derecho laboral, se modificó el Estatuto de los Trabajadores por la Ley 12/2021, de 28 de septiembre (conocida como *Ley Rider*)⁵²⁸, para reconocer el derecho a la información de los representantes de los trabajadores acerca del

⁵²⁶ Ap. 2.5.2.1 B).

⁵²⁷ Estos conceptos pueden variar en función del contexto pertinente e incluir el entendimiento de la correcta aplicación de los elementos técnicos durante la fase de desarrollo del sistema de IA, las medidas que deben aplicarse durante su uso, las formas adecuadas de interpretar la información de salida del sistema de IA y, en el caso de las personas afectadas, los conocimientos *necesarios para comprender el modo en que las decisiones adoptadas con la ayuda de la IA tendrán repercusiones para ellas*.

⁵²⁸ Ley 12/2021, de 28 de septiembre, por la que se modifica el texto refundido de la Ley del Estatuto de los Trabajadores, aprobado por el Real Decreto Legislativo 2/2015, de 23 de octubre, para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales («BOE» núm. 233, de 29 de septiembre de 2021).

uso de algoritmos en la empresa. Se introduce una nueva letra d) en el artículo 64.4 («*Derechos de información y consulta y competencias*»), donde se establece que «el comité de empresa, con la periodicidad que proceda en cada caso, *tendrá derecho a ser informado por la empresa de los parámetros, reglas e instrucciones en los que se basan los algoritmos o sistemas de inteligencia artificial que afectan a la toma de decisiones que pueden incidir en las condiciones de trabajo, el acceso y mantenimiento del empleo, incluida la elaboración de perfiles*».

Este derecho de información colectiva se suma al derecho de información individual ya reconocido en el RGPD. Pero debe destacarse una importante diferencia entre ambos, ya que, mientras el derecho de información individual se reconoce exclusivamente en caso de decisiones *plena o íntegramente* automatizadas que afecten al trabajador, el nuevo derecho de información colectiva (o más bien, la obligación de la empresa), opera también cuando se utilicen algoritmos *como simple apoyo a la toma de decisiones por parte de la empresa*. Dicho en otras palabras, aunque el algoritmo no sea determinante para la decisión final tomada sobre la persona trabajadora, su mero uso, implica el nacimiento de los derechos de información de la representación legal de la plantilla.

En consecuencia, todas las empresas –tanto plataformas digitales como empresas tradicionales– que tengan personas asalariadas a su cargo y utilicen algoritmos o sistemas de IA, *tanto plenamente automatizados como semiautomatizados*, para la gestión de personas (incluyendo decisiones de selección, contratación, asignación de horarios, tareas, medición de productividad, promociones, fijación de salarios, despido, etc.) tienen obligación de dar información algorítmica.

El contenido de la información que la empresa debe proporcionar a la representación legal de la plantilla ha de comprender «los parámetros, reglas e instrucciones en los que se basan los algoritmos o sistemas de inteligencia artificial». Sin perjuicio de la falta de consenso técnico sobre la interpretación de estos términos, dicha referencia debe interpretarse como la obligación de la empresa de proporcionar información sobre las variables y los parámetros o importancia relativa de cada variable en el algoritmo, y las reglas de programación que conducen a la toma de la decisión⁵²⁹.

⁵²⁹ Para mayor concreción sobre esa obligación de información en el ámbito laboral, puede verse el documento del MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL, *Información algorítmica en el ámbito laboral. Guía Práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral*, mayo 2022, donde se concretan y sistematizan las obligaciones empresariales de información, con el fin de garantizar el derecho fundamental a la no discriminación, entre otros. Se diferencian cuatro apartados: *información general, información sobre la lógica y funcionamiento del sistema, información sobre las consecuencias, y otra información relevante*. Cada epígrafe consta de una serie de preguntas como, por ejemplo, qué tipo de tecnología utiliza el algoritmo (caja negra, aprendizaje continuo) quién lo ha desarrollado, con qué decisiones, si hay personas en este proceso de decisión y en qué medida participan, qué datos de entrenamiento se han utilizado, si se han verificado

A pesar de la existencia de estos dos derechos de información individual y colectiva, la doctrina iuslaboralista considera que siguen siendo insuficientes para abordar los retos jurídicos que plantea la opacidad algorítmica por varias razones: por un lado, porque el derecho de información colectiva reconocido por el art. 64.4 ET se refiere únicamente a los parámetros y métricas utilizadas por el algoritmo, pero no a los efectos que el algoritmo genera, siendo esta información muy útil a la hora de evaluar su impacto discriminatorio, especialmente su impacto indirectamente discriminatorio⁵³⁰; por otro lado, porque el RGPD consagra el derecho a conocer los datos personales que recopila la empresa pero no las inferencias a las que ha llegado el sistema a partir de los datos disponibles o el perfil que ha elaborado⁵³¹; por último, porque el derecho de información sobre la lógica del algoritmo reconocido en los arts. 13, 14 y 15 RGPD se refiere sólo a decisiones íntegramente automatizadas pero excluye aquellas en las que también hay intervención humana.

La última limitación apuntada ha sido superada por la consagración en el art. 86 RIA del *derecho a una explicación* que también opera en los casos de decisiones semiautomatizadas. Además, por lo que se refiere concretamente al ámbito laboral, el art. 26.7 RIA impone a los empleadores que utilicen sistemas de IA de alto riesgo en el lugar de trabajo la *obligación de informar tanto a los representantes de los trabajadores como a los trabajadores afectados de que estarán expuestos al mismo, con carácter previo a su uso o puesta en servicio*⁵³². Dicho derecho de informa-

los extremos, qué consecuencias pueden derivarse para las personas la decisión adoptada, si se ha informado a la persona candidata del uso de algoritmos.

⁵³⁰ GINÉS I FABRELLAS, A., *op.cit.*, p. 319. Sobre este nuevo derecho de información colectiva reconocido en el art. 64.4 la autora señala que, si bien se trata de una regulación pionera a nivel europeo que permite a la representación legal de la plantilla valorar la legalidad –e, incluso, potencial discriminatorio– de las variables utilizadas por el algoritmo, debería haberse acompañado con información estadística sobre el impacto del uso de dicha tecnología.

⁵³¹ No obstante, en la Guía Práctica del MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL sobre *Información algorítmica en el ámbito laboral*, cit., se mantiene que, en caso de elaboración algorítmica de perfiles, la empresa debería dar información sobre los tipos de perfiles que hace el algoritmo y concretamente a la persona trabajadora se le ha de informar sobre la asignación concreta a uno de ellos.

⁵³² Esta obligación del empleador se establece como parte de las obligaciones de los responsables del despliegue de los sistemas de IA de alto riesgo recogidas en el art. 26. En concreto, su apartado 7 dispone que «Antes de poner en servicio o utilizar un sistema de IA de alto riesgo en el lugar de trabajo, los responsables del despliegue que sean empleadores informarán a los representantes de los trabajadores y a los trabajadores afectados de que estarán expuestos a la utilización del sistema de IA de alto riesgo. Esta información se facilitará, cuando proceda, con arreglo a las normas y procedimientos establecidos en el Derecho nacional y de la Unión y conforme a las prácticas en materia de información a los trabajadores y sus representantes». También *vid.*, el Considerando 92 donde se establece que se deberá velar por que se informe a los trabajadores y a sus representantes sobre el despliegue de estos sistemas en el lugar de trabajo incluso aunque no se cumplan las condiciones de las obligaciones de información o de información y consulta previstas en otros instrumentos jurídicos.

ción se considera *accesorio* a otros que puedan tener los trabajadores en virtud del Derecho o las prácticas nacionales o de la Unión (a los que no debe afectar), y *necesario* para el objetivo de protección de los derechos fundamentales que subyace en el Reglamento⁵³³. También se establece en el art. 2, relativo al ámbito de aplicación, que lo establecido en el RIA no impedirá que la Unión o los Estados miembros mantengan o introduzcan disposiciones legales, reglamentarias o administrativas más favorables para los trabajadores en lo que atañe a la protección de sus derechos respecto al uso de sistemas de IA por parte de los empleadores, ni que fomenten o permitan la aplicación de convenios colectivos más favorables para ellos.

Otra norma laboral con la que se trata de hacer frente a la opacidad algorítmica y fomentar la transparencia de la IA es la *Ley 3/2023, de 28 de febrero, de Empleo*. Dicha norma establece la creación de herramientas de apoyo a la toma de decisiones basadas en evidencias estadísticas para la mejora de la empleabilidad de las personas⁵³⁴. En concreto, es el art. 17 el que se ocupa de la «toma de decisiones fundamentada en el análisis de datos, las evidencias estadísticas y el análisis del mercado de trabajo».

Dicho precepto prevé, en primer lugar, que para el diseño de dichas herramientas se utilizará la información contenida en el Sistema Público Integrado de Información de los Servicios de Empleo, que se desarrollarán en el seno del Sistema Nacional de Empleo, con la colaboración de las Comunidades Autónomas, y que se pondrán a disposición de las personas responsables de la tutorización y seguimiento individualizado de las personas demandantes de los servicios de empleo o de las personas responsables de la prestación de servicios a personas, empresas o entidades empleadoras usuarias, para facilitar su toma de decisiones y las recomendaciones que elaboren y emitan en el ejercicio de sus funciones.

A continuación, se establece que en la toma de estas decisiones y recomendaciones se actuará siempre en defensa de los derechos al empleo y al acceso a los servicios públicos de empleo de la persona interesada y, por lo que se refiere al derecho a la igualdad y no discriminación, se evitará la discriminación directa o indirecta. Para conseguir dicho objetivo, el siguiente inciso del apartado segundo establece que se evaluarán anualmente los efectos de las mismas para

⁵³³ Dispone el Considerando 92 que «El presente Reglamento se entiende sin perjuicio de la obligación de los empleadores de informar o de informar y consultar a los trabajadores o a sus representantes, en virtud del Derecho o las prácticas nacionales o de la Unión, incluida la Directiva 2002/14/CE del Parlamento Europeo y del Consejo, por la que se establece un marco general relativo a la información y a la consulta de los trabajadores⁴⁰, sobre la decisión de poner en servicio o utilizar sistemas de IA».

⁵³⁴ El art. 36.3 establece que «el uso de herramientas de apoyo a la toma de decisiones basadas en el análisis de datos, en las evidencias estadísticas y en el análisis del mercado de trabajo, contribuirá también a identificar en cada momento qué personas o colectivos de personas están encontrando mayores dificultades para acceder al empleo [...]».

producir las modificaciones que sean necesarias para un tratamiento igualitario y no discriminatorio de cada individuo demandante de empleo.

Se prevén dos garantías importantes para los afectados por las decisiones y recomendaciones basadas en el análisis de datos y en las evidencias estadísticas: por un lado, que las mismas siempre podrán ser revisadas o modificadas por el personal responsable de la tutorización y seguimiento individualizado de cada persona demandante de servicios; y por otro lado, que los usuarios tendrán garantizado el acceso al contenido íntegro de estas decisiones y recomendaciones. Por tanto, nos encontramos ante otro derecho de información individual del que sería titular el demandante de ocupación que interactúa con el servicio público de empleo.

Por último, se prevé un desarrollo reglamentario del conjunto de instrucciones que compondrán el algoritmo subyacente a las decisiones y recomendaciones basadas en el análisis de datos y las evidencias estadísticas, y se prohíbe que estas herramientas puedan utilizar datos personales pertenecientes a las categorías especiales reguladas en el art. 9.1 RGPD.

2.5.5.2. *Derecho público español*

A) *Acceso a la información algorítmica*

El interés por garantizar una IA transparente se acentúa en el ámbito público, donde el principio y mandato de transparencia desempeña un papel fundamental en los Estados democráticos. De hecho, está reconocido en el art.105.b) de nuestra Constitución y su desarrollo legislativo se llevó a cabo en un primer momento por la *Ley 30/1992, de 26 de noviembre, de Régimen Jurídico de las Administraciones Públicas y del Procedimiento Administrativo Común (LRJPAC)* –en cuyo art. 37 se recogía el derecho de los ciudadanos a acceder a los «archivos y registros administrativos»–, y posteriormente, por la *Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno*, y las leyes autonómicas que la han desarrollado, cuyo objeto es ampliar y reforzar la transparencia de la actividad pública, regular y garantizar el derecho de acceso a la información relativa a aquella actividad y establecer las obligaciones de buen gobierno que deben cumplir los responsables públicos así como las consecuencias derivadas de su incumplimiento (art. 1).

Dicho principio de transparencia implica que todas las personas tienen derecho a acceder a la información pública en los términos del art.105.b) CE y de la Ley de Transparencia, a participar en los asuntos públicos y, sobre todo, a controlar a la Administración supervisando cómo se toman las decisiones.

La creciente automatización de la actividad administrativa y el uso de algoritmos para dictar resoluciones dificulta la transparencia y trazabilidad necesarias para garantizar la fiscalización de las decisiones públicas y deberían implicar, como contrapartida, la posibilidad de ejercitar un derecho de acceso al algoritmo. El Consejo de Transparencia y Buen Gobierno (en adelante, CTBG) en su resolución R/0058/2021⁵³⁵, de 20 de mayo, se pronunció sobre la trascendencia de la IA en el ámbito público y la necesidad de garantizar la transparencia algorítmica en los siguientes términos: *«en el contexto actual de progresivo desarrollo e implantación la administración electrónica y uso creciente de la inteligencia artificial, los algoritmos están adquiriendo una relevancia decisiva, a la vez que se incrementa su complejidad. Pueden sustentar la toma de decisiones públicas o, directamente, ser fuente de decisiones automatizadas con consecuencias muy relevantes para las personas. Esta evolución está generando una creciente demanda ciudadana de transparencia de los algoritmos utilizados por las administraciones públicas como condición inexcusable para preservar la rendición de cuentas y la fiscalización de las decisiones de los poderes públicos y, en último término, como garantía efectiva frente a la arbitrariedad o los sesgos discriminatorios en la toma de decisiones total o parcialmente automatizadas»*. Se añadía, a continuación, que *«[m]ientras no se instauren otros mecanismos que permitan alcanzar los fines señalados con garantías equivalentes –como podrían ser, por ejemplo, auditorías independientes u órganos de supervisión–, el único recurso eficaz a tales efectos es el acceso al algoritmo propiamente dicho, a su código, para su fiscalización, tanto por quienes se puedan sentir perjudicados por sus resultados como por la ciudadanía en general, en aras de la observancia de principios éticos y de justicia»*. Por tanto, dicha autoridad administrativa aboga en este caso por facilitar el acceso al código fuente de los algoritmos para luchar contra las decisiones injustas y discriminatorias, posicionamiento que se reitera en las resoluciones RT/253/2021, de 19 de noviembre y RT/748/2021, de 10 enero 2022, estimándose en ambos casos las reclamaciones relativas al acceso al código fuente de la aplicación informática utilizada para el sorteo de tribunales asociado a procesos selectivos en materia educativa en la Comunidad de Madrid⁵³⁶.

⁵³⁵ file:///C:/Users/ASUS/Downloads/R-0058-2021-bis.pdf

⁵³⁶ Dichas resoluciones suponen un cambio de criterio respecto al mantenido en su Resolución 701/2018, de 18 de febrero de 2019 (Caso del algoritmo Bosco para la concesión del bono social eléctrico) donde excluía expresamente al código fuente de la obligatoriedad de su cesión en aplicación de la LTBG. Por lo que respecta a la RT/253/2021, fue recurrida por la Comunidad de Madrid y el recurso contencioso-administrativo fue estimado por la Sentencia n° 158/2022 del Juzgado Central Contencioso-Administrativo n° 11, reconociendo que las solicitudes de información fueron debidamente atendidas por dicha Comunidad sin que procediera efectuar actuación alguna adicional por parte de la misma. El juzgado confirma la aplicación del límite del artículo 14.1.j) de la LTAIBG alegado por la Comunidad para denegar el acceso al código fuente y considera que la petición del solicitante puede considerarse abusiva, vinculando el abuso a su carácter excesivo e injustificado, que

A mayor abundamiento, en el caso de que la Administración emplee sistemas de IA para adoptar decisiones, la doctrina defiende la existencia de obligaciones específicas de publicidad activa en virtud de las cuales se debería informar a los ciudadanos no sólo del *uso de algoritmos*, sino también del *tipo de algoritmo elegido en el caso concreto* y proporcionar *información* suficiente sobre el mismo que permita a los interesados conocer los *motivos* que subyacen en la adopción de la decisión⁵³⁷. Para HUERGO LORA, «en la medida en que un algoritmo sirva para tomar decisiones que afecten a un ciudadano, la Administración debe hacerlo público del mismo modo que cualquier otro documento que forme parte del expediente administrativo, en el que el algoritmo se integraría en virtud del artículo 70.1 LPAC puesto que “sirve de antecedente y fundamento a la resolución administrativa” [...]. El interesado tendrá acceso al expediente, en todo caso, si interpone recurso contencioso-administrativo (artículo 48 LJ), pero también podrá acceder sin necesidad de llegar al proceso judicial, como interesado en el procedimiento administrativo [artículo 53.1. a) LPAC]». Además, en opinión de este autor, que el algoritmo haya sido diseñado por un tercero (contratista de la Administración) y pueda estar protegido como secreto empresarial, no autoriza a la Administración a denegar el acceso a la información⁵³⁸.

excedería del contenido razonable del derecho de acceso a una información pública cabalmente considerado, teniendo en cuenta que el administrado tuvo conocimiento, en todo caso, del funcionamiento del sistema para la designación de los vocales de los tribunales de procesos selectivos cuyo resultado y funcionamiento es público. Por tanto, concluye que en el caso concreto se había dado suficiente transparencia al procedimiento público y que los interesados tuvieron la posibilidad de acudir al sorteo público celebrado, teniendo acceso a los resultados.

⁵³⁷ Por todos, HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, pp. 85 y ss. También sobre la publicidad activa y el derecho de acceso a la información puede verse, TIMÓN HERRERO, M., «La transparencia de la actuación de las Administraciones Públicas. Cinco desafíos y algunas propuestas», recuperado de <https://elderecho.com/transparencia-actuacion-de-administraciones-publicas>. Para LAZCOZ MORATINOS / CASTILLO PARRILLA, J.A., *op.cit.*, p. 221, cuando los algoritmos se utilizan en el sector público, su opacidad es aún más incomprensible debido a que actúan funcionalmente como normas, por lo que no se entiende el frontal incumplimiento de un principio tan básico como el de publicidad normativa. Y para SUBERBIOLA GARBIZU, I., *op.cit.*, tanto si se trata de algoritmos que adoptan sus decisiones en *black boxes*, como si el contenido del algoritmo y su código fuente no se publica, o simplemente es ininteligible, el resultado puede ser el mismo, esto es, la anulación o en su caso la anulabilidad del acto administrativo. En la misma línea, VESTRI, G., *op.cit.*, p. 382, considera necesario verificar la transparencia tanto en la elección del algoritmo por parte de la Administración, como en su propio funcionamiento, garantizando la accesibilidad del interesado a conocer cómo el algoritmo adopta la decisión. De esta manera, se pretende alcanzar un doble nivel de transparencia que es trascendental precisamente en el seno de una herramienta intangible como es el algoritmo.

⁵³⁸ HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, p. 86. En tales casos, mantiene este autor, «la Administración contratante deberá adoptar en el pliego las cláusulas necesarias para poder después enseñar a terceros los elementos del modelo o de la aplicación necesarios para hacer visible y controlable su funcionamiento. Si no lo hace, es decir, si se encuentra después en un conflicto entre el cumplimiento del contrato y el respeto a los derechos del interesado en el

De hecho, algunas leyes de transparencia autonómicas ya recogen deberes de información en relación con el uso de algoritmos. Así, la pionera *Ley 1/2022, de 13 de abril, de Transparencia y Buen Gobierno de la Comunidad Valenciana*, en su art. 16 (letra l) dispone que las Administraciones Públicas deberán publicar: «*La relación de sistemas algorítmicos o de inteligencia artificial que tengan impacto en los procedimientos administrativos o la prestación de los servicios públicos con la descripción de manera comprensible de su diseño y funcionamiento, el nivel de riesgo que implican y el punto de contacto al que poder dirigirse en cada caso, de acuerdo con los principios de transparencia y explicabilidad*».

También la *Ley 15/2022 integral de igualdad*, en su art. 23, establece el deber de las administraciones públicas de poner en marcha mecanismos para que los algoritmos involucrados en la toma de decisiones tengan en cuenta criterios de *minimización de sesgos, transparencia y rendición de cuentas*, y que se priorice la *transparencia en el diseño y la implementación*, así como la *capacidad de interpretación de las decisiones adoptadas por los mismos*, precepto que, no obstante su relevancia a la hora de sentar parámetros generales, ha sido considerado por la doctrina como una mera declaración de intenciones que deberá concretarse en medidas específicas a cumplir por las administraciones y las empresas⁵³⁹.

Por otro lado, y aunque no tiene carácter vinculante, la Carta de Derechos Digitales reconoce a los ciudadanos una serie de derechos frente a la Administración cuando se utilicen sistemas inteligentes en la toma de decisiones, que se traducen en cargas o deberes de información y transparencia para los entes públicos. Así, se recogen en su apartado XVIII.6 los derechos a:

procedimiento, su obligación es, claramente, respetar estos últimos, aunque ello tenga, en su caso, costes indemnizatorios frente al contratista por incumplimiento del contrato. Estos costes serían la consecuencia de una defectuosa configuración del contrato».

Sobre el hecho de que el autor sea una persona física o jurídica con la que ha contratado la Administración, señala el CTBG en su RT/253/2021, de 19 de noviembre, que lo habitual es que en estos casos la primera ceda los derechos de propiedad intelectual en favor de la segunda. De hecho, el artículo 308 de la Ley 9/2017, de 8 de noviembre, de Contratos del Sector Público, establece lo siguiente: «1. Salvo que se disponga otra cosa en los pliegos de cláusulas administrativas o en el documento contractual, los contratos de servicios que tengan por objeto el desarrollo y la puesta a disposición de productos protegidos por un derecho de propiedad intelectual o industrial llevarán aparejada la cesión de este a la Administración contratante. En todo caso, y aun cuando se excluya la cesión de los derechos de propiedad intelectual, el órgano de contratación podrá siempre autorizar el uso del correspondiente producto a los entes, organismos y entidades pertenecientes al sector público».

⁵³⁹ Mantiene COTINO HUESO, L., «La primera sentencia del Tribunal de Justicia de la Unión Europea...», *cit.*, que más allá de elementos generales, hay que actualizar la legislación de transparencia para imponer y concretar la publicidad activa en la forma de registros públicos de algoritmos. Deben determinarse unos mínimos de información a facilitar en estos registros y en algunos casos incluso podría exigirse la apertura de algoritmos para examinar su funcionamiento y resultado.

a) Que las decisiones y actividades en el entorno digital respeten los *principios de buen gobierno y el derecho a una buena Administración digital*, así como los *principios éticos que guían el diseño y los usos de la inteligencia artificial*. b) La *transparencia sobre el uso de instrumentos de inteligencia artificial y sobre su funcionamiento y alcance en cada procedimiento concreto* y, en particular, acerca de los *datos utilizados, su margen de error, su ámbito de aplicación y su carácter decisorio o no decisorio*. La ley podrá regular las condiciones de transparencia y el acceso al código fuente, especialmente con objeto de verificar que no produce resultados discriminatorios. c) Obtener una *motivación comprensible en lenguaje natural* de las decisiones que se adopten en el entorno digital, con *justificación de las normas jurídicas relevantes, tecnología empleada*, así como de los *criterios de aplicación de las mismas al caso*. El interesado tendrá derecho a que se motive o se explique la decisión administrativa cuando esta se separe del criterio propuesto por un sistema automatizado o inteligente. d) Que la adopción de decisiones discrecionales quede reservada a personas, salvo que normativamente se prevea la adopción de decisiones automatizadas con garantías adecuadas». Y en el siguiente apartado se recoge la necesidad de llevar a cabo *evaluaciones de impacto* en los derechos digitales a la hora de diseñar los algoritmos cuando se adopten decisiones automatizadas o semiautomatizadas.

Además de este marco de referencia que ofrece la Carta para garantizar los derechos de los ciudadanos en la era digital, el Gobierno español ha creado la Agencia de Supervisión de la IA (AESIA) entre cuyas funciones principales está la de minimizar los riesgos para la integridad, la intimidad, la igualdad de trato y la no discriminación, en particular entre mujeres y hombres, y demás derechos fundamentales que pueden verse afectados por el mal uso de los sistemas de IA (salud, seguridad, etc.)⁵⁴⁰. Con dichos fines, este organismo pretende dotar de transparencia y auditabilidad a los tratamientos masivos de datos personales basados en IA y velar porque se cumpla la regulación europea en materia de IA⁵⁴¹.

⁵⁴⁰ <https://portal.mineco.gob.es/es-es/comunicacion/Paginas/agencia-espa%C3%B1ola-de-supervisi%C3%B3n-de-inteligencia-artificial.aspx>. Vid., el Real Decreto 729/2023, de 22 de agosto, por el que se aprueba el Estatuto de la Agencia Española de Supervisión de Inteligencia Artificial. Como pone de relieve la Estrategia española sobre IA 2024, la AESIA, será un actor clave a la hora de definir las buenas prácticas en el uso de la IA y de garantizar la evaluación de los modelos que se vayan desarrollando. Tiene la competencia de asegurar que en España se despliegan sistemas de IA seguros, responsables y éticos. Vid., pp. 54-55.

⁵⁴¹ Por lo que se refiere en concreto a los sesgos discriminatorios, y con el objetivo de evitarlos y reforzar la protección de los usuarios, se establece entre los fines de dicha Agencia, «el fomento de entornos reales de prueba de los sistemas de IA» (art. 4.3.d). Y entre sus competencias se incluye, en el art. 10.1.c), «El apoyo al desarrollo y uso de sistemas de inteligencia artificial con perspectiva de género, incorporando el principio de igualdad de oportunidades entre mujeres y hombres en su diseño y ejecución y promoviendo la realización de evaluaciones de impacto capaces de identificar po-

En otros países, como Reino Unido, se han elaborado estándares de transparencia para los algorítmicos públicos cuya finalidad es servir de marco de referencia (no vinculante) a las organizaciones del sector público para que proporcionen información clara a los ciudadanos sobre las herramientas algorítmicas que utilizan y los motivos de su utilización. La publicación de este estándar por la Oficina Central Digital y de Datos del Reino Unido (CCDO), vino motivada por la recomendación del Centro de Ética e Innovación de Datos (CDEI) que, tras realizar un estudio sobre el sesgo en la toma de decisiones basadas en algoritmos, aconsejó al Gobierno imponer una obligación de transparencia obligatoria a las organizaciones del sector público que utilizaran algoritmos para respaldar decisiones importantes que afectaran a las personas. Dicho estándar de transparencia tiene dos niveles: en el primero se incluye una *breve descripción del algoritmo, incluyendo cómo, por qué y para qué se está utilizando*, mientras que el segundo incluye información más detallada sobre *cómo funciona la herramienta, los conjuntos de datos que se han utilizado para entrenar el modelo y el nivel supervisión o vigilancia humana*⁵⁴².

B) Motivación de las resoluciones

Con carácter general, el deber de motivación de las resoluciones (del tipo que sean) tiene un triple objetivo: por un lado, evitar arbitrariedades de los responsables de su adopción; por otro lado, dar a conocer a los interesados las razones o argumentos por las que se adopta la decisión para que, si lo consideran conveniente, puedan rebatirlas; y por último, apreciar si se ha actuado o no dentro de los límites legales en la toma de decisiones⁵⁴³.

La opacidad algorítmica puede dificultar el cumplimiento de dicho deber poniendo en riesgo los derechos reconocidos en el CEDH y en la CDFUE sobre las garantías de obtención de una resolución judicial fundada en derecho

sibles sesgos discriminatorios por cualquiera de los motivos prohibidos por el ordenamiento jurídico. Todo ello con el objetivo de eliminar los sesgos discriminatorios de cualquier tipo y, en particular, los sesgos de género y los de índole étnico-racial».

⁵⁴² Con más detalle, EGUÍLUZ CASTAÑEIRA, J.U., «Reino Unido avanza en la transparencia de los algoritmos», <https://www.cuatrecasas.com/es/spain/propiedad-intelectual/art/internacional-reino-unido-avanza-en-la-transparencia-de-los-algoritmos>

⁵⁴³ Precisamente en esta línea se dirigieron los alegatos esgrimidos por los abogados de la Fundación Civio en su recurso contra la sentencia de 30 de diciembre de 2021 del Juzgado Central de lo Contencioso-Administrativo número 8, en la que se denegaba el acceso al código fuente del algoritmo *Bosco* empleado por la Administración para decidir quienes tenían derecho a recibir el bono social eléctrico. En dicho recurso se mantiene que «impedir el acceso al código fuente de los programas implica la imposibilidad de que el ciudadano pueda verificar si una herramienta a través de la cual se le aplica el derecho se halla acorde con la ley o incumple la misma». El recurso puede consultarse en: <https://civio.app.box.com/s/8514had075ntq3o7tymfx8fq0jxcxgmf>

(esto es, el derecho a obtener una sentencia motivada inherente al derecho a un proceso equitativo –art. 6 CEDH– que se recoge taxativamente en el art. 45 CEDH), así como al derecho a la tutela judicial efectiva del art. 47 CDFUE. Pero también vulnera el derecho a una buena administración recogido en el art. 41 CDFUE, cuyo apartado 2.c) establece «la obligación que incumbe a la administración de motivar sus decisiones», deber que deriva del art. 296 del Tratado de Funcionamiento de la Unión⁵⁴⁴.

Por lo que se refiere a nuestra Carta Magna, el TS tiene sentado que el derecho a la tutela judicial efectiva consagrado en el art. 24 CE *«incluye el derecho a obtener de los órganos judiciales una respuesta razonada, motivada y congruente con las pretensiones oportunamente deducidas por las partes»*⁵⁴⁵. Por tanto, la motivación de las sentencias es un derecho fundamental y su falta conduce a la arbitrariedad. En igual sentido, el artículo 120.3 CE determina expresamente que *«las sentencias serán siempre motivadas y se pronunciarán en audiencia pública»*. Este deber de motivación consiste, como apunta la STS de 25 de junio de 2015, *«en la exteriorización del iter decisorio o conjunto de consideraciones racionales que justifican el fallo. De esta forma, la motivación de las sentencias se presenta como una exigencia constitucional establecida en el artículo 120.3 CE configurándose como un deber inherente al ejercicio de la función jurisdiccional en íntima conexión con el derecho a la tutela judicial efectiva que establece el artículo 24 CE (STC 144/2003 de julio y STS de 5 de diciembre de 2009)»*. Y por lo que se refiere a los actos administrativos, su motivación se conecta igualmente con el derecho fundamental a la tutela efectiva y la prohibición de indefensión (STS de 12 de mayo de 1999 y 25 de junio de 1999). Por ende, el administrado debe conocer el fundamento, circunstancias o motivos del acto que le afecta y debe realizarse con la amplitud necesaria para su debido conocimiento y posterior defensa⁵⁴⁶.

⁵⁴⁴ La doctrina viene manteniendo que el derecho a obtener una resolución motivada forma parte del derecho a una buena administración. En este sentido, RIBES RIBES, A., «La posición del contribuyente ante los sistemas de inteligencia artificial utilizados por la Administración», *Revista Aranzadi*, 2021; o MARÍN GARCÍA, S., «Ética e inteligencia artificial». Cuadernos de la Cátedra CaixaBank, 2019. Para SUBERBIOLA GARBIZU, I., *op.cit.*, el art. 41 CEDF se imbrica directamente con los objetivos que persigue la utilización de la inteligencia artificial por parte de los entes públicos, como es el derecho a una buena administración. Sobre este derecho señala PONCE SOLÉ, J., «A propósito de la sentencia de la Audiencia Nacional de 30 de abril de 2024, núm. de rec. 51/2022, sobre el caso BOSCO», que del mismo deriva una obligación de diligencia o cuidado que da lugar a un examen del expediente, la decisión y su motivación para hallar no sólo errores sino omisiones o deficiencias (por ejemplo, no tomar en consideración elementos relevantes o no hacerlo con el debido cuidado). Recuperado de <https://almacenderecho.org/por-que-se-equivocan-las-sentencias-sobre-el-algoritmo-del-bono-social-electrico>

⁵⁴⁵ Es doctrina constitucional, entre otras, sentada en las SSTC 108/2001, de 23 de abril y 68/2011, de 16 de mayo.

⁵⁴⁶ El art. 35 de la Ley 39/2015, del Procedimiento Administrativo Común de las Administraciones Públicas se refiere en general a la motivación de los actos administrativos.

Siendo esto así, cualquier resolución emanada de un poder público (a través de sistemas de IA o al margen de ellos) debería respetar estas exigencias, teniendo derecho el interesado a una explicación de los motivos de tales decisiones, lo que no siempre será fácil cuando se emplean sistemas de IA para su adopción.

La falta de motivación o la motivación defectuosa constituye un defecto formal, pero las consecuencias dependen de la importancia del defecto. En virtud del art. 48.2 de la *Ley 39/2015, de 1 de octubre, del Procedimiento Administrativo Común de las Administraciones Públicas* (en adelante, LPAC), el defecto de forma sólo determinará la anulabilidad cuando el acto carezca de los requisitos formales indispensables para alcanzar su fin o dé lugar a la indefensión de los interesados⁵⁴⁷. En consecuencia, lo relevante jurídicamente es que la motivación sea *suficiente*. Así, según señala el TS, *basta con que la motivación sea breve y sucinta, pero, en todo caso, tiene que ser suficiente*⁵⁴⁸.

Cuando se emplean sistemas de IA para la toma de decisiones públicas, se ha mantenido que éstas han de ser lo suficientemente motivadas como para conocer cuál ha sido el algoritmo utilizado, de modo que omitirlo y no dar información sobre el mismo provocaría su invalidez. La motivación de estos actos consistiría en la explicación transparente, en el propio acto o en algún otro medio a disposición del interesado, del proceso lógico que conduce a su adopción. Se trataría de poner a disposición de los interesados del acto administrativo cuál ha sido la programación del algoritmo para que estos puedan verificar su adecuación a Derecho y conocer si, por ejemplo, se han adoptado

⁵⁴⁷ GAMERO CASADO, E., «Necesidad de motivación e invalidez de los actos administrativos sustentados en inteligencia artificial o en algoritmos», en <https://almacenderecho.org/necesidad-de-motivacion-e-invalidez-de-los-actos-administrativos-sustentados-en-inteligencia-artificial-o-en-algoritmos>, recuerda que «la motivación es una exigencia meramente formal: necesidad de que en el propio acto administrativo consten las razones por las que se dicta. No se trata, por tanto, de que existan motivos para dictar el acto sino de que esos concretos motivos sean transparentes y visibles, que se hagan constar en el propio acto. No constituye una garantía material o de fondo, sino formal, aunque la exigencia de motivar se establezca para dar soporte a un derecho material como el de no padecer indefensión».

⁵⁴⁸ STS de 15 de diciembre de 1999. Sobre la motivación de los actos administrativos señala también el TS que «tiene un carácter finalístico que consiste en impedir “que el interesado se vea privado de los medios de defensa necesarios para impugnar la actuación de la Administración” (STS de 7 de octubre de 1998). Además, según la jurisprudencia contencioso-administrativa, los actos administrativos carentes de motivación (cuando ésta es preceptiva) generan indefensión (SSTS 1198/2019, de 19 de septiembre y 198/2019, de 19 de febrero, entre otras muchas). Es más, cuando se trata de derechos y libertades fundamentales, la motivación tiene alcance constitucional (SSTC 26/1981, 236/2007, 17/2009 y 46/2014, entre otras muchas), por lo que es causa de nulidad de pleno Derecho del acto administrativo en virtud del art. 47.1 a) LPAC.

criterios discriminatorios o si los mismos son objetivos y no se incurre en la arbitrariedad proscrita por los art. 9.1 y 103.2 CE⁵⁴⁹.

Se observa, pues, que la motivación en este tipo de actos está íntimamente conectada con la transparencia. La doctrina sostiene que, a pesar de que la transparencia (o publicidad) y la motivación son elementos distintos, su régimen jurídico es distinto y el impacto de la IA en cada uno de esos ámbitos también lo es, existe un punto de conexión entre el funcionamiento opaco de los algoritmos, la falta de transparencia y la deficiente (o nula) motivación de la actuación administrativa⁵⁵⁰. En este sentido, «si en un contexto de elevada complejidad un decisor público se ayuda de un algoritmo para tomar la decisión, y es muy difícil conocer cómo ha llegado el programa informático al resultado que ofrece al decisor público, difícilmente se podrá motivar esa decisión de forma adecuada, ya que la motivación podrá ir poco más allá de indicar vagamente que la decisión adoptada es conforme al resultado obtenido con la aplicación del programa. Tampoco se daría buena cuenta de las razones y el razonamiento empleado con disponer de las instrucciones de programación si tenemos en cuenta elementos de autoaprendizaje del algoritmo (como el *machine o deep learning*), puesto que el funcionamiento del programa puede haber cambiado, sin poderse determinar cuánto ni en qué elementos»⁵⁵¹.

Las decisiones públicas han de ser objeto de especial justificación en términos que sean comprensibles para el ciudadano. No sería suficiente con que la motivación fuera comprensible por otra máquina, ni que la misma se li-

⁵⁴⁹ Así, GAMERO CASADO, E., *op.cit.*, quien explica que la sentencia del Consejo de Estado italiano, núm.8472 de 13 de diciembre de 2019 exige, como requisito para la aplicación de algoritmos, el «pleno conocimiento del módulo utilizado y de los criterios aplicados», así como que «este conocimiento del algoritmo debe garantizarse en todos los aspectos: desde sus autores hasta el procedimiento utilizado para su elaboración, el mecanismo de decisión, las prioridades asignadas en el procedimiento de evaluación y toma de decisiones y los datos seleccionados como relevantes (...) a fin de poder verificar que los criterios, condiciones y resultados del procedimiento robótico cumplen con las prescripciones y las finalidades establecidas por la ley o por la propia administración sobre dicho procedimiento y para que queden claras –y, por lo tanto, sean cuestionables– las modalidades y reglas a partir de las que se haya programado».

⁵⁵⁰ CAPDEFERRO VILLAGRASA, O., «La inteligencia artificial en el sector público: desarrollo y regulación de la actuación administrativa inteligente en la cuarta revolución industrial», *Revista de internet, derecho y política*, n.º 30, Marzo, 2020, p. 10. Para RIBES RIBES, A. *op.cit.*, «conviene recordar que el artículo 3.1.c) LRJSP se refiere a la transparencia como principio general que debe presidir la actuación administrativa. Estrechamente ligado a este principio, aunque sin ser coincidentes, aparece el concepto de la motivación, o lo que es igual, el derecho del contribuyente a obtener una decisión (automatizada) motivada. Este derecho forma parte del derecho a una buena administración y, asimismo, deriva del principio de interdicción de la arbitrariedad administrativa, consagrado a nivel constitucional».

⁵⁵¹ CAPDEFERRO VILLAGRASA, O., *op.cit.loc.cit.* Esto lleva a que un sector doctrinal mantenga que, en tanto en cuanto los actos administrativos basados en algoritmos no puedan motivarse y su funcionamiento pueda explicarse, no deberíamos emplearlos. Por todos, ALCOLEA AZCÁRRAGA, C., *op.cit.*

mitara a indicar de forma vaga que la decisión finalmente adoptada por el humano es conforme al resultado alcanzado con la aplicación del programa. Además, en los casos de decisiones discrecionales en los que el sistema de IA se utilizara como apoyo a la toma de decisión, si ésta se apartara del resultado ofrecido por el sistema, el mandato del artículo 35.1.i) LPAC obligaría a explicar las razones por las cuales se ignoró la solución algorítmica⁵⁵².

En definitiva, si bien es lícito que a través de la IA se pretenda aumentar la eficiencia de la Administración, en ningún caso dicho objetivo debería alcanzarse a costa de mermar la motivación de los actos administrativos o de que la Administración fuera menos transparente.

2.5.6. Pronunciamientos judiciales y de órganos administrativos en materia de transparencia y acceso a la información algorítmica

En los últimos años, diversos tribunales y órganos administrativos han tenido que resolver asuntos relacionados con el uso de algoritmos y su opacidad. Mientras que en EEUU han sido más reacios con carácter general a reconocer un derecho de acceso a información sobre el algoritmo basándose en la necesidad de proteger los derechos de propiedad intelectual o el secreto empresarial⁵⁵³, en Europa, algunos pronunciamientos se muestran proclives a garantizar un nivel de transparencia mucho mayor y obligaciones de información más exigentes en los casos en que se utilicen sistemas de IA para la toma

⁵⁵² Concluye con rotundidad GAMERO CASADO, E., *op.cit.*, que aun cuando no haya un precepto que exija concretamente motivar los actos administrativos que se basan en algoritmos o IA, «tal requisito es inequívocamente predicable en los supuestos establecidos con carácter general por el artículo 35.1 LPAC».

⁵⁵³ Importante ejemplo de lo afirmado lo constituye la sentencia del caso COMPAS (*State v. Loomis*), ya analizada anteriormente, donde el tribunal en ningún momento cuestiona la protección del algoritmo de COMPAS como secreto empresarial imponiendo al recurrente el deber de demostrar la incorrección de los resultados, produciéndose una verdadera inversión de la carga de la prueba en tanto que se exige al acusado que demuestre que los resultados de COMPAS son incorrectos (sin permitirle, además, acceder a las fórmulas de sus algoritmos), mientras que ni la acusación –ni Northpointe– están obligados siquiera a explicar el funcionamiento del sistema, su metodología y la fiabilidad de sus evaluaciones, las cuales se presumen válidas a pesar de las dudas suscitadas por algunos estudios. Sobre el posicionamiento de los tribunales estadounidenses respecto al acceso al código fuente de los sistemas predictivos que se emplean concretamente en el ámbito de la justicia, explica SOLAR CAYON, *op.cit.*, pp. 139 y ss, que los tribunales penales han venido denegando con carácter general dicho acceso a los afectados en base a un *privilege procesal*, reconocido en la jurisprudencia civil, que exceptúa *prima facie* a las compañías propietarias de la obligación de revelar el código fuente de sus sistemas. Esto supone una carga muy gravosa para el acusado porque para tener acceso al mismo tendrá que probar que esa información es relevante y necesaria para su defensa, que no tener acceso a ella le causaría un daño específico y que no la puede obtener por ninguna otra vía.

decisiones, principalmente en el ámbito público, aunque esta tendencia no se encuentra suficientemente consolidada y también encontramos resoluciones en sentido contrario.

Entre las resoluciones que se decantan por la transparencia, cobra particular relevancia la Sentencia del Tribunal de la Haya de 5 de febrero de 2020, caso *SyRI*, que será objeto de análisis a continuación⁵⁵⁴. En sentido contrario, se denegó el acceso al código fuente del algoritmo a los trabajadores de Uber en Reino Unido en la sentencia del Tribunal de Ámsterdam C/13/692003/HA RK 20-302, de 11 de marzo de 2021, y a los de Glovo en Italia en el caso resuelto por la sentencia del Tribunal Ordinario de Bolonia de 27 de noviembre de 2020⁵⁵⁵. Dentro de nuestras fronteras destaca el caso del algoritmo *BOSCO* respecto del cual ha recaído en fechas muy recientes una nueva sentencia (esta vez de la Audiencia Nacional) que desestima el recurso contra la sentencia 143/2021 del Juzgado Central de lo Contencioso-Administrativo, de 30 de diciembre, que a su vez desestimaba el recurso contra la decisión del CTBG frente a la petición de acceso al código fuente del sistema utilizado para la concesión del bono social eléctrico, del que también daremos buena cuenta en las siguientes páginas. Al hilo de este caso se pondrá de manifiesto que la postura de dicho Consejo estatal sobre el acceso al código fuente del algoritmo tampoco es uniforme, existiendo resoluciones a favor y en contra.

⁵⁵⁴ Además de esta sentencia, también se obligó a dar acceso al código fuente de diferentes sistemas automatizados empleados por el sector público en los siguientes asuntos: Tribunal Regional Administrativo del Lazio-Roma, Sección III bis, Sentencias N° 3769, de 22 de marzo de 2017 y N° 10964, de 13 de septiembre de 2019; Comisión Francesa de Acceso a los Documentos Administrativos, Decisiones N° 20144578, de 8 de enero de 2015 y N° 20180276, de 19 de abril de 2018. En nuestro país, la Resolución de 21 de septiembre de 2016 de la Comisión de Garantía del Derecho de Acceso a la Información Pública (GAIP), que estima las Reclamaciones 123/2016 i 124/2016 (acumuladas) sobre el algoritmo matemático que determina la selección de los miembros de los tribunales correctores de las pruebas de acceso a la Universidad (PAU) y DNI de los candidatos presentados, y su Resolución 200/2017, de 21 de junio relacionada con las anteriores. Por lo que se refiere al CTBG, las resoluciones R/0058/2021 de 20 de mayo, sobre el acceso al algoritmo para calcular pensiones de la TGSS; RT/253/2021 de 19 de noviembre y RT/ 748/2021 de 10 de enero 2022 sobre acceso al código fuente de una aplicación utilizada para el sorteo de tribunales asociados a procesos selectivos para educación en Madrid, y la Resolución del Expediente 551-2023 sobre el acceso a información del Sistema de Seguimiento Integral de los casos de Violencia de Género (Sistema VioGén).

⁵⁵⁵ En el primer caso, el Tribunal de Ámsterdam en sentencia C/13/692003/HA RK 20-302 de 11/3/2021 (y C/13/687315/HA RK 20-207 de la misma fecha) resuelve desfavorablemente, en tanto que se proporcionó a los trabajadores una explicación suficiente sobre dicho funcionamiento y esta no fue impugnada por ellos (sin perjuicio de la obligación de proporcionar acceso a los datos personales a los afectados). El segundo fue resuelto por sentencia del Tribunal ordinario de Bolonia de 27/11/2020, que estimó que los algoritmos utilizados por la empresa no eran inclusivos. http://www.era-comm.eu/oldoku/SNLLaw/14_Other_topics/122DV68f_RIVAS_VALLEJO_Paper_ES.pdf

2.5.6.1. *Dos casos paradigmáticos (y antagónicos) sobre transparencia algorítmica en el sector público: el caso SyRi y el caso Bosco.*

A) *Caso SyRi*

Especial atención y relevancia se otorga a la transparencia algorítmica en la sentencia del Tribunal de la Haya de 5 de febrero de 2020, que resuelve el conocido caso *SyRi* (acrónimo de *System Risk Indication*). Se trataba de un sistema de análisis o evaluación del riesgo utilizado por el Gobierno holandés para prevenir y combatir posibles fraudes a la Seguridad Social y a Hacienda por parte de los solicitantes de beneficios estatales. La aplicación, a disposición de los ayuntamientos de los municipios holandeses, procesaba una enorme cantidad de datos personales de los contribuyentes (entre otros, ingresos, pensiones, seguros, tipo de casa, impuestos, multas, integración, educación, deudas o subsidio de desempleo, antecedentes penales...) para calcular después mediante algoritmos quién tenía más probabilidades de defraudar a la Administración. Tras combinar esas amplias categorías de datos (previamente almacenados por separado) y analizarlos utilizando un modelo de riesgo (no revelado), el programa arrojaba una serie de conclusiones sobre el nivel de riesgo de las personas de cometer fraude fiscal, que posteriormente era evaluado por una unidad que fijaba su nivel de riesgo definitivo.

A pesar de que el Tribunal reconoce que el objetivo perseguido por el Gobierno holandés era legítimo y de gran relevancia social, puesto que pretendía prevenir y combatir el fraude en interés del bienestar económico (párrafo 6.4), considera que *el algoritmo no respetaba los principios de proporcionalidad y transparencia de la Administración en la utilización de una aplicación basada en IA*, añadiendo que el modelo de riesgo elaborado por SyRi podía tener efectos no deseados, como *estigmatizar y discriminar* a la ciudadanía, no sólo por la ingente cantidad de información que recogía, sino también porque sólo se utilizaba en barrios considerados problemáticos en los que residían personas con rentas bajas o minorías⁵⁵⁶.

La argumentación de la sentencia se construye en base a la vulneración del derecho al respeto de la vida privada y familiar consagrado en el art. 8 CEDH y no, como cabía esperar, en base a la vulneración del RGPD, y más concretamente, de su art. 22⁵⁵⁷. Tiene en consideración los principios sobre

⁵⁵⁶ Párrafo 6.92. Para MERCADER UGUINA, J.R., *op.cit.*, p. 8, esta sentencia recuerda a la STJUE (Gran Sala), de 16 de julio de 2015, en el asunto C-83/14, *Chez Razpredelenie Bulgaria AD*, en la que se debatía la proporcionalidad de una medida que suponía un perjuicio excesivo para un grupo étnico que se veía estigmatizado sobre la base de consideraciones puramente económicas.

⁵⁵⁷ Los comentaristas de esta sentencia se lamentan de que se haya dejado pasar la oportunidad para aclarar el contenido de dicho derecho y justifican el posicionamiento del tribunal en

los que se asienta la normativa de protección de datos, esto es, la transparencia, limitación de la finalidad y minimización de datos, para considerarlos propios del art. 8.2 CEDH. Por tanto, partiendo de la protección del derecho al respeto de la vida privada, en el que se incluye el derecho a la protección de los datos personales, el tribunal considera que «todos deben poder tener la expectativa razonable de que, en el caso de uso de SyRI, se respetará suficientemente su vida privada, algo que, lamentablemente, no se puede garantizar». Concluye que SyRI *no es lo suficientemente transparente y verificable*.

Destaca la sentencia que en el uso de las tecnologías emergentes existe una responsabilidad especial para encontrar un *equilibrio justo y adecuado* entre las ventajas que aportan estos sistemas predictivos para la prevención y el control del fraude, y la injerencia que puedan suponer en el derecho al respeto de la vida privada. Consideró que no se cumplía esta exigencia de proporcionalidad puesto que la aplicación permitía el acceso a una gran cantidad de datos de la persona, hasta el punto de que llegaba a vulnerar su vida privada, lo cual resultaba desproporcionado pues el interés social que se pretendía obtener (evitar el fraude) no era comparable a la invasión a la privacidad que se producía⁵⁵⁸. Concluye que el empleo de SyRI vulnera el art. 8 CEDH tanto

que la vulneración del derecho a la vida privada planteaba una vía mucho más clara para establecer la ilegalidad de la norma. Se trata del primer procedimiento que utiliza como fundamento jurídico el incumplimiento de un Derecho Humano para limitar el uso de determinadas tecnologías, abriendo una nueva línea argumental jurídica para futuros casos. Vid., <https://www.thetechnolawgist.com/2020/02/12/derechos-humanos-en-un-mundo-de-algoritmos-la-sentencia-historica-que-ata-en-corto-la-implantacion-de-modelos-opacos/>

⁵⁵⁸ SyRI realizaba una recogida masiva de datos que no cumplía con el principio de minimización de datos del art.5.c) RGPD. El párrafo 4.17 de la sentencia recoge los tipos de datos que pueden analizarse por el sistema en base a su normativa y la sentencia reconoce que dichas categorías reflejan casi toda la información relativa a una persona. Considera el tribunal que la lista de categorías elegibles es tan exhaustiva que prácticamente no hay datos personales que no puedan ser objeto de tratamiento por SyRI (párrafo 6.98). Como explica SUBERBIOLA GARBIZU, I., *op.cit.*, a juicio del Tribunal, la herramienta no cumple el requisito establecido en el art. 8.2 CEDH de que la injerencia en el ejercicio del derecho al respeto de la vida privada debe ser necesaria, proporcionada y subsidiaria en relación con su propósito previsto. Para llegar a esta conclusión contrasta el contenido de SyRI a la luz de los propósitos que persigue, con la invasión de la privacidad que su uso implica, y opina que la legislación no cumple con el «equilibrio justo» (la relación razonable) que debe existir en el CEDH, entre el interés social al que sirve la legislación, y la violación de la vida privada que crea, para poder hablar sobre una invasión de la privacidad suficientemente justificada. En el mismo sentido, BATTAGLINI MANRIQUE DE LARA «Sentencia histórica del Tribunal de la Haya anulando la elaboración de perfiles para el fraude de la Seguridad Social (SyRI)» <https://www.linkedin.com/pulse/sentencia-hist%C3%B3rica-del-tribunal-de-la-haya-anulando-manuela/>; LAZCOZ MORATINOS/ CASTILLO PARRILLA, *op.cit.* p. 218; COTINO HUESO, L., «Holanda: “SyRI, ¿a quién sanciono?” Garantías frente al uso de inteligencia artificial y decisiones automatizadas en el sector público y la sentencia holandesa de febrero de 2020», *La Ley privacidad*, N.º. 4 (abril-junio 2020).

por la falta de transparencia del modelo de riesgo empleado, como por su uso sesgado, ya que sólo se usaba en determinados barrios.

Por otro lado, es destacable que a pesar de que el sistema estaba rodeado de garantías (por ejemplo, anonimización de datos, división funcional, borrado y confidencialidad) el tribunal las considera insuficientes por no cumplir con el requisito de la transparencia. En particular, echa en falta auditorías del algoritmo y considera que hubiera sido deseable realizar la evaluación de impacto prevista en el art. 35.10 RGPD⁵⁵⁹. Por tanto, la carencia más grave que aprecia el tribunal, y que le lleva a considerar que las muchas garantías previstas en este caso fueron insuficientes, fue la opacidad del algoritmo y la falta de información a los interesados cuyos datos se tratan.

Por lo que se refiere la opacidad del sistema, la sentencia afirma que el modelo de riesgo utilizado y los indicadores de riesgo que constituyen este modelo de riesgo son «secretos» y desconocidos por los interesados (párrafo 6.65). El sistema no proporciona información sobre cómo determinados datos o circunstancias pueden derivar en el incremento del riesgo (párrafo 6.87), ni sobre el funcionamiento del modelo de riesgo, el tipo de algoritmo, ni el método de análisis de riesgos aplicado (párrafo 6.89). El Gobierno justifica la falta de información alegando que pretendía evitar prácticas o tecnología inversa por posibles defraudadores que les permitiera eludir su calificación como perfil de riesgo a inspeccionar y que sólo siendo oscuro podrían obtenerse datos masivos del comportamiento de los ciudadanos de la suficiente calidad (párrafo 6.49)⁵⁶⁰.

Los jueces señalaron que no habían recibido suficiente información sobre el sistema ya que el Estado no había revelado la forma de calcular la probabilidad o riesgo de fraude. Por tanto, aunque la Administración hubiera emplea-

⁵⁵⁹ Apuntan LAZCOZ MORATINOS / CASTILLO PARRILLA, *op.cit.*, p. 221, que aunque se había hecho una evaluación de impacto de la normativa previa a su implementación, el tribunal considera que dicha evaluación debía realizarse obligatoriamente por cada proyecto SyRI que se desarrollase (párrafo 6.105). El GT 29 pone de relieve la importancia de las evaluaciones de impacto de protección de datos del art.35 RGPD como instrumento a la hora de paliar los efectos discriminatorios. Por ello, concluyen los autores citados, que en el ámbito de la toma de decisiones automatizada del art. 22 RGPD, habría que considerar obligatoria dicha evaluación de forma previa a la implementación de estos algoritmos.

⁵⁶⁰ HUERGO LORA, A., «Una aproximación a los algoritmos...», *cit.*, p. 86, habla del *gaming* (engaño del algoritmo) para referirse al peligro de que la revelación de su contenido puede mermar su eficacia. Es decir, si los ciudadanos saben cuáles son los criterios que utiliza el algoritmo para calificarles, ajustarán su conducta a esos criterios para «engañar» al algoritmo. Se ha utilizado este argumento para denegar el acceso al código fuente en casos como el del algoritmo Bosco. Así, vid., el *Informe técnico sobre la oportunidad de la entrega del código fuente de la aplicación informática para la comprobación de los requisitos exigidos a los solicitantes del bono social*, de 4/12/2019.

do un algoritmo de caja blanca, la negativa a proporcionar información sobre el mismo conculcaba igualmente los derechos de los administrados, ya que, en palabras del tribunal «es difícil comprender cómo un interesado podría defenderse contra el hecho de que se haya presentado un informe de riesgos sobre él o ella» (párrafo 6.90).

Respecto al derecho de información de los interesados sobre el tratamiento de sus datos personales, se advierte que la normativa SyRi no establece la obligación de informar individualmente o notificar «puerta a puerta» de que se ha realizado un informe de riesgos (párrafo 6.53). Además, quienes no son considerados perfil de riesgo, no pueden saber cómo se procesaron sus datos. Que deban destruirse los datos a las cuatro semanas, «no altera el requisito de transparencia con respecto a ese procesamiento» (párrafo 6.89). Los interesados «tampoco son informados automáticamente después», sino que sólo se les informa si hay un control e investigación en respuesta a un informe de riesgos (párrafo 6.53).

Finalmente, por lo que se refiere al derecho a la igualdad y la no discriminación (extremo que más nos interesa en este trabajo) la sentencia también considera que la transparencia, es necesaria frente a los peligros de «exclusión o discriminación injustificadas». De hecho, los demandantes y el Relator Especial de la ONU consideraron que SyRi tenía un «efecto discriminatorio y estigmatizador», pues al investigar los vecindarios más marginados contribuía a los estereotipos y reforzaba una imagen negativa de los ocupantes de dichos vecindarios. Aunque el tribunal entendió que el uso en ciertos barrios no implicaba por sí solo una desproporción en los términos del art. 8.2 del CEDH, consideró que el procesamiento de grandes cantidades de datos (entre los que se incluían datos de especial protección) sí constituía un riesgo de discriminación en función del nivel socioeconómico o el origen migrante (párrafo 6.93).

B) *Caso Bosco*

Especial mención merece también el caso del programa *BOSCO* originado dentro de nuestras fronteras⁵⁶¹. Se trata de un *software*, de titularidad pública pero gestionado por empresas privadas, que determina quién tiene acceso al bono social eléctrico, es decir, qué consumidores en situación de vulnerabilidad tienen derecho a un descuento en la factura de la luz.

⁵⁶¹ Sobre este caso, puede verse, entre otros, BAZ LOMBA, C., «Los algoritmos y la toma de decisiones administrativas. Especial referencia a la transparencia», *Revista CEFLegal*, 243, 2021, pp. 119-160; también COTINO HUESO, L., «Caso Bosco, a la tercera tampoco va la vencida. Mal camino en el acceso los algoritmos públicos», *Diario LA LEY*, n° 84, Sección Ciberderecho, 17 de Mayo de 2024.

La concesión de este beneficio no es una decisión discrecional de la Administración, sino que las condiciones para obtenerlo y la forma de solicitarlo están previstas en la ley⁵⁶². Además, es la propia Administración la que proporciona a las empresas privadas la información sobre el consumidor, por tanto, no hay margen para que la empresa deniegue el bono a alguien que reúna los requisitos reglamentarios⁵⁶³. No obstante, ante las abundantes quejas de usuarios a los que se les había denegado la ayuda y creían tener derecho ella, la Fundación CIVIO solicitó a la Administración, al amparo de la Ley de Transparencia, que le proporcionara su especificación técnica, los resultados de las pruebas de comprobación de la aplicación y su código fuente, a fin de comprobar si el programa presentaba algún defecto o errores en su configuración que impidiera la correcta aplicación de la legislación⁵⁶⁴.

Frente a la desestimación de la solicitud por silencio administrativo del Ministerio para la Transición Ecológica, CIVIO acudió al CTBG que estimó parcialmente su reclamación y obligó al Gobierno a entregar parte de la documentación solicitada para poder comprobar si el sistema adolecía de defectos de diseño. En concreto, se dio acceso a la especificación técnica de la aplicación y el resultado de las pruebas realizadas para comprobar que cumplía su especificación funcional. Se consideró que dicha información no quedaba protegida por el derecho de propiedad intelectual contemplado en la Directiva que los ampara. Por lo que se refiere a las primeras, el Consejo manifestaba que «pueden incluir aspectos, entre otros, como si es un sistema operativo de código abierto, cómo realiza el almacenamiento de datos, cuál es su lenguaje de programación o si incluye herramientas para depuración de memoria y análisis del rendimiento del software. Existen multitud de especificaciones técnicas de programas de ordenador expuestas al público en Internet»; y en relación a las segundas, «no inciden en el *hardware* protegido y sirven a nuestro juicio al propósito perseguido por la LTAIBG de controlar la acción pública y el proceso de toma de decisiones»⁵⁶⁵.

⁵⁶² Se regula en el Real Decreto 897/2017 (desarrollado por la Orden ETU/943/2017) que aplica una tarifa eléctrica reducida a consumidores vulnerables.

⁵⁶³ DE LA NUEZ SÁNCHEZ CASCADO, E., «Algoritmos y transparencia», disponible en <https://www.hayderecho.com/2020/02/18/algoritmos-y-transparencia-2/>

⁵⁶⁴ En principio, la aplicación debía ser neutral porque solo tenía que realizar la mera aplicación de los requisitos de acceso a la prestación sin cambiar nada del panorama legal ni práctico. Sin embargo, tras su implantación, la mitad de los anteriores beneficiarios dejaron de ser idóneos para el reconocimiento del bono social. De aquí se infiere, tal como explica RIVAS VALLEJO, P., «Sesgos de automatización...», *cit.*, pp. 52-53, que aunque en principio las reglas del juego sean idénticas, la forma de trabajar de los algoritmos cambia los resultados porque eleva exponencialmente el índice de granularidad del análisis de datos disponibles.

⁵⁶⁵ Seguía diciendo esta Resolución que «En modo alguno, conocer las especificaciones técnicas de una aplicación informática implica, ni directa ni indirectamente, acceder a los datos persona-

Sin perjuicio de lo anterior, negó el acceso al código fuente alegando que quedaba protegido por derechos de propiedad intelectual, y en concreto, por el límite previsto en el art. 14. j) de la Ley de Transparencia, referente a la propiedad intelectual de los programas informáticos⁵⁶⁶. Sobre esta cuestión, el CTBG hace las siguientes consideraciones:

«Este límite se relaciona con la propia naturaleza de lo solicitado: el código fuente que sirve de soporte a la aplicación telemática que permite al comercializador de referencia comprobar que el solicitante del bono social cumple los requisitos para ser considerado consumidor vulnerable. El código fuente es el archivo o conjunto de archivos que tienen un conjunto de instrucciones muy precisas, basadas en un lenguaje de programación, que se utiliza para poder compilar los diferentes programas informáticos que lo utilizan y se puedan ejecutar sin mayores problemas. Los usuarios pueden usar el *software* sin mayores preocupaciones gracias a una interfaz gráfica sencilla que se basa en el desarrollo del código fuente. El usuario no necesita saber el lenguaje de programación utilizado para desarrollar un determinado *software*. El *software* ha sido extraordinariamente difícil de clasificar como materia específica de propiedad intelectual debido a que su doble naturaleza plantea problemas particulares para quienes tratan de establecer analogías con las categorías jurídicas existentes. Esta es la razón por la que ha habido intentos de clasificarlo como objeto de derechos de autor, de patentes o de secretos comerciales, e incluso como un derecho *sui generis* de *software*»⁵⁶⁷.

Tras las anteriores apreciaciones, concluye que «[p]uesto que el código fuente se expresa de forma escrita, resulta lógico pensar que el *software* puede ser protegido por el derecho de autor como obra literaria», argumento que la Fundación rebate al entender que las obras realizadas en las Administraciones

les que después se vayan a incluir en la misma. El resultado de las pruebas, que también se solicita, no debe incluir, en ningún caso, datos personales reales, sino simplemente estadísticos sobre la eficacia del sistema en función de la finalidad que se persigue».

⁵⁶⁶ En dicho precepto se establece que el «El derecho de acceso podrá ser limitado cuando acceder a la información suponga un perjuicio para: j) *El secreto profesional y la propiedad intelectual e industrial*».

⁵⁶⁷ Sigue diciendo el Consejo que este es «el enfoque vigente respecto de la protección del *software* en diversos tratados internacionales. Así, por ejemplo, el artículo 4 del Tratado de la OMPI sobre Derecho de Autor (WCT), el artículo 10 del Acuerdo sobre los ADPIC de la Organización Mundial del Comercio y el artículo 1 de la Directiva (91/250/CEE) del Consejo Europeo sobre la protección jurídica de programas de ordenador equiparan el software con las obras literarias, protegidas por el derecho de autor. A efectos de la Directiva, el término «programa de ordenador» incluye programas en cualquier forma, incluso los que están incorporados en el «hardware»; este término designa también el trabajo preparatorio de concepción que conduce al desarrollo de un programa de ordenador, siempre que la naturaleza del trabajo preparatorio sea tal que más tarde pueda originar un programa de ordenador».

Públicas no están sujetas a los derechos de autor, y que lo contrario podría suponer la denegación de cualquier documento elaborado por un trabajador público⁵⁶⁸.

Ante dicha negativa, CIVIO presentó un recurso contencioso-administrativo que fue desestimado por el Juzgado Central de lo Contencioso-Administrativo número 8 en sentencia de 30 de diciembre de 2021⁵⁶⁹, basándose en que el acceso al código fuente afectaría no sólo a los derechos de propiedad intelectual sobre el *software*, sino también a la seguridad pública y a la defensa nacional por la conexión del sistema con bases de datos de carácter sensible cuya integridad podía verse comprometida, aceptando con ello todos los argumentos alegados por el Gobierno. Además, según estableció la sentencia (en la línea mantenida por el Abogado del Estado), lo que hace el programa informático no es un acto administrativo en sí mismo (lo que reforzaría la pretensión de CIVIO para acceder al código fuente bajo el principio de legalidad para conocer cómo toma una u otra decisión, frente a la protección de la propiedad intelectual del código) porque no es la aplicación quien decide, sino un órgano administrativo⁵⁷⁰.

Esta sentencia fue recurrida en apelación por CIVIO y nuevamente se desestimaron sus pretensiones⁵⁷¹. La Sala de lo Contencioso-Administrativo de la Audiencia Nacional (sección séptima), en su reciente sentencia de 30 de abril de 2024, rechaza por tercera vez la petición de acceso al código fuente empleando para ello argumentos muy similares a los ya expuestos: por un lado, el límite del art. 14.1.j) LTIBG; y por otro lado, razones de seguridad pública⁵⁷².

⁵⁶⁸ También se ha alegado por un sector doctrinal que, en la medida en que en los casos de sistemas de IA que no tienen margen de decisión porque se limitan a aplicar la normativa al afecto, el código haría las veces de «norma», esto es, el código sería ley, sería inaplicable el límite de acceso basado en la propiedad intelectual. En este sentido, por todos, PONCE SOLÉ, J., «A propósito de la sentencia de la Audiencia Nacional de 30 de abril de 2024, núm. de rec. 51/2022, sobre el caso BOSCO».

⁵⁶⁹ JUR\2022\137045

⁵⁷⁰ El Juzgado repuso que la Administración «ajusta su actuación a dicha normativa, dictando el correspondiente acto administrativo. Y para ello utiliza una aplicación informática, denominada “sistema de información BOSCO”, que se inserta en una fase del procedimiento administrativo, cuyo objeto es verificar el cumplimiento de los requisitos establecidos previamente por la normativa citada. Siendo lo anterior así, no puede considerarse que el acto administrativo se dicte por una aplicación informática, sino por un órgano administrativo, y en caso de que el destinatario de dicho acto esté disconforme con el mismo, podrá impugnarlo en vía administrativa, y en vía judicial. Por tanto, la legalidad del acto administrativo no está justificada por la aplicación informática que instrumentalmente se utiliza en una fase del correspondiente procedimiento administrativo, sino por la normativa que regula la materia».

⁵⁷¹ El recurso puede consultarse en <https://civio.app.box.com/s/8514had075ntq3o7tymfx8fq0jcxgmf>

⁵⁷² En opinión de COTINO HUESO, L., «Caso Bosco, a la tercera tampoco va la vencida...», cit., «lo que se alega –y triunfa– por la puerta de atrás es la seguridad informática, que no había sido alegada convenientemente por la Administración cuando tocaba». Los tribunales se basan en diversos

Reitera que el código fuente está protegido por la Ley de Propiedad Intelectual y que excluir a la Administración del derecho de la protección intelectual carece de todo fundamento⁵⁷³. Además, considera que la difusión del código fuente de la aplicación BOSCO pondría en grave riesgo derechos de terceros y atentaría a bienes jurídicos protegidos por los límites al derecho de acceso a la información pública del artículo 14.1 letras d), g), i) y k). Concluye que la ocultación del código fuente de la aplicación es, sin duda alguna, garantía de protección ante las vulnerabilidades.

Y respecto a la necesidad de transparencia para asegurar las garantías del art. 9.3 CE viene a reiterar lo argumentado en la sentencia recurrida, esto es, que «la aplicación informática no es más que una herramienta que se integra en el seno de un procedimiento administrativo que seguirá produciendo un acto administrativo plenamente sujeto al ordenamiento jurídico, de modo que el control de la legalidad de los actos administrativos sigue estando garantizado: *la legalidad (o no) del acto administrativo no viene justificada por el uso de la aplicación informática sino por la adecuación al ordenamiento jurídico del acto producido*».

Pues bien, como puede observarse sin dificultad y de forma similar a lo ocurrido en el caso SyRi, la piedra angular aquí también es la transparencia algorítmica. Sobre esta cuestión, advertía el abogado de CIVIO que aceptar la interpretación del CTBG permitirá a la Administración pública desarrollar algoritmos ocultos, sosteniendo que un Estado social y democrático de Derecho jamás debería permitir que se nos regulara mediante códigos fuente o algoritmos secretos⁵⁷⁴. Además, se planteaba la cuestión del contenido y extensión de la información que debía darse al ciudadano.

No obstante, debe apuntarse que, a diferencia del caso SyRi, en el caso Bosco el algoritmo no tenía margen para decidir la concesión o denegación de la ayuda social porque se trataba de una aplicación reglada. HUERGO LORA hace una serie de apreciaciones interesantes sobre este punto: en primer lugar, que las exigencias de transparencia y motivación no tienen que ser idénticas en los casos en que se utilicen sistemas de IA que influyen en el contenido de la decisión administrativa al no existir un marco normativo que lleve a una única

Informes técnicos para mantener que revelar el código fuente podría comprometer la seguridad de la información y aumentar la vulnerabilidad ante ataques informáticos, lo que podría afectar adversamente la integridad de datos personales y la seguridad pública.

⁵⁷³ «El artículo 7.2 de la Ley 33/2003 de Patrimonio de las Administraciones Públicas atribuye la consideración de patrimoniales de la Administración a los derechos de propiedad incorporal hayan sido adquiridos de particulares o generados por la propia Administración. Y el artículo 157 de la Ley 40/2015, de Régimen Jurídico del Sector Público se refiere expresamente a los derechos de propiedad intelectual de que son titulares las Administraciones Públicas sobre las aplicaciones desarrolladas por sus servicios o que hayan sido objeto de contratación».

⁵⁷⁴ Así finalizaban los fundamentos materiales de la demanda de Civio.

decisión jurídicamente admisible, y los casos en que se utilicen aplicaciones «regladas» como en el caso Bosco. En estos últimos, los errores podrían detectarse comparando lo que dice la aplicación (denegación del bono, por ejemplo) con lo que resulta del marco normativo (que puede ser su concesión); en segundo lugar, que si bien sería deseable conocer todo el funcionamiento interno de la Administración (transparencia total), dicho grado de transparencia no sería necesario para poder corregir errores o ilegalidades en actuaciones regladas y tampoco se exige a los humanos cuando son ellos los que toman las decisiones sin apoyo informático (a pesar de que el ‘código fuente’ de los humanos, es decir, lo que motiva su comportamiento, es infinitamente más complejo e inaccesible que el de un ordenador). Además, no podría decirse que este tipo de aplicaciones informáticas «deciden» sobre los derechos de los ciudadanos, en la medida en que lo único que hacen es aplicar un marco normativo que sólo admite una solución correcta y están sujetas a recurso cuando no lo cumplen⁵⁷⁵.

La postura de los distintos órganos decisores en este caso ha sido objeto de críticas por entender la doctrina que la solicitud de transparencia de la Fundación CIVIO estaba más que justificada, ya que no puede tolerarse que se adopten decisiones desestimatorias de beneficios aprobados por el Gobierno en favor de determinados colectivos sin conocer cómo se llega a ellas. Esto generaría indefensión y un enorme riesgo de arbitrariedad de los entes públicos. Como señala DE LA NUEZ SÁNCHEZ-CASCADO, «no se trata de proteger la propiedad intelectual de un programa –más allá de las disquisiciones que realiza el CTBG acerca de si el software puede o no estar protegido por el derecho de autor como obra literaria– para que nadie lo copie o lo plagie. Se trata sencillamente de determinar qué es lo que puede estar funcionando mal en la configuración técnica del mismo para que se esté denegando una ayuda

⁵⁷⁵ HUERGO LORA, A., «Por qué aciertan las sentencias sobre el ‘algoritmo’ del bono social eléctrico», *Almacén del Derecho*, 10 de mayo de 2024. Insiste el autor en que cuando las solicitudes se gestionan sin aplicación informática nadie pretende tener acceso a la mente del funcionario para ver cómo puede cometer errores a la hora de aplicar la normativa. Si se deniega la prestación a alguien que tiene derecho a ella, se presenta un recurso y, si se comprueba el error, se estima y se le reconoce el derecho, sin necesidad de indagar si el error se produjo por inadvertencia, por sobrecarga de trabajo o porque había entendido mal la normativa. El hecho de que se utilice un apoyo informático no debería cambiar las cosas. Esta última idea se pone de relieve en la sentencia 158/2022 del Juzgado Central de lo Contencioso-Administrativo, de 31 de octubre de 2022: «En cualquier decisión administrativa, hay casi siempre un reducto inaccesible de valoración e indeterminación en el órgano o funcionario resolutor, legitimando su labor por la presunción de objetividad e imparcialidad en su actuación. Tampoco debemos olvidar que siempre queda el recurso a la tutela judicial efectiva que tendría el administrado caso de que entendiera que la actuación administrativa –con la utilización de la aplicación informática– no resultara imparcial o independiente, con la aportación de casos concretos que evidenciaran que el sistema funciona de modo incorrecto o discriminatorio. En definitiva, el sistema implementado estaría siempre sujeto a un control *ex post*, y no necesariamente *ex ante*, que por otra parte suele ser lo usual en los distintos ámbitos sometidos a control judicial».

a gente que sí reúne los requisitos para acceder a ella. En todo caso, parece que en este caso el test del daño y el test del interés público que siempre hay que realizar cuando se pretende aplicar uno de los límites del art. 14 de la Ley 19/2013 debería decantarse claramente a favor de la transparencia»⁵⁷⁶.

Para SOLÉ PONCE, los riesgos aducidos en la sentencia de la Audiencia Nacional conllevarían que pudiera denegarse en cualquier caso el acceso al código fuente, postura que se separa de la que están adoptando otros tribunales de nuestro entorno y órganos administrativos como el CTBG⁵⁷⁷. Sin perjuicio de que, además, las distintas sentencias recaídas en este caso concreto obvian cuestiones relevantes como los posibles errores del programa que llevarían a una multiplicación a gran escala de denegaciones erróneas (afectando además a colectivos vulnerables), la aprobación del uso del sistema algorítmico por la Administración sin participación ciudadana ni publicidad, o el protagonismo de las empresas privadas del sector (comercializadores de referencia), que parece conducir a una privatización de la actividad procedimental administrativa, al encargar a éstos, con un simple reglamento, el desarrollo de funciones administrativas como comprobar, denegar o aprobar y notificar.

Éste no es el único caso sobre decisiones algorítmicas que ha sido objeto de resolución judicial por nuestros tribunales y órganos administrativos, aunque sí el más relevante y polémico. También la STS 1231/2020, de 1 de octubre, resolvió un caso en el que se había tomado la decisión de entrar en un domicilio en base al perfil de alto riesgo de un sujeto determinado por un algoritmo⁵⁷⁸.

Por último, tanto el CTBG estatal como la GAIP, se han pronunciado en diversas ocasiones sobre el acceso al código fuente llegando la mayo-

⁵⁷⁶ Para esta autora, lo mínimo que se debería exigir es conocer cómo funciona el sistema algorítmico y cómo adopta decisiones que son de enorme relevancia para personas en situación o riesgo de pobreza energética. Criticando también esta postura, COTINO HUESO, L., «Caso Bosco...» *cit.*

⁵⁷⁷ PONCE SOLÉ, J., «A propósito de la sentencia de la Audiencia Nacional de 30 de abril de 2024...», *cit.*

⁵⁷⁸ ECLI: ES:TS:2020. En esta sentencia se dice lo siguiente: «No pueden servir de base, para autorizar la entrada, los datos o informaciones generales o indefinidos procedentes de estadísticas, cálculos o, en general, de la comparación de la situación supuesta del titular del domicilio con la de otros indeterminados contribuyentes o grupos de estos, o con la media de sectores de actividad en todo el territorio nacional, sin especificación o segmentación detallada alguna que avale la seriedad de tales fuentes. Tal análisis, de hacerse excepcionalmente, debe atender a todas las circunstancias concurrentes y, muy en particular, a que de tales indicios, vestigios o datos generales y relativos –verificado su origen, seriedad y la situación concreta del interesado respecto a ellos– sea rigurosamente necesaria la entrada, lo que exige valorar la existencia de otros factores circunstanciales y, en particular, la conducta previa del titular en respuesta a actuaciones o requerimientos de información efectuados por la Administración».

ría de las veces (aunque no todas) a soluciones favorables al interesado que después han sido anuladas por los órganos jurisdiccionales⁵⁷⁹. Así ocurrió, por ejemplo, con la resolución RT/253/2021 que fue recurrida por la Comunidad de Madrid y el recurso contencioso-administrativo fue estimado por la Sentencia n° 158/2022 del Juzgado Central Contencioso-Administrativo n° 11.

2.5.7. Otras propuestas para hacer frente a la opacidad algorítmica

Ante la inexistencia de normas que obligaran a las empresas a dar acceso al código fuente de sus algoritmos o a divulgar sus características ante la reclamación de un particular, la doctrina y algunos grupos de expertos también propusieron soluciones para garantizar la transparencia algorítmica.

En el ámbito del Derecho público, dichas propuestas han ido desde la necesidad de que las empresas que desarrollen y vendan estos sistemas a las Administraciones Públicas deban renunciar a sus derechos de propiedad intelectual y a cualquier otro privilegio legal que impida la plena comprensión y control del *software*, hasta la necesidad de exigir que sean las propias Administraciones las que desarrollen sistemas algorítmicos de código abierto⁵⁸⁰. Pero en general, se aboga por soluciones intermedias que, en la línea seguida en el contexto europeo, procuren conseguir un equilibrio entre la transparencia algorítmica y los intereses empresariales. En este sentido, se viene manteniendo que para conseguir una IA confiable habría que informar de las capacidades y limitaciones del sistema de IA, de aquello que se puede esperar de él de forma realista y del seguimiento de los protocolos de seguridad en cuanto a su implementación, sin perjuicio de que el usuario deba saber con

⁵⁷⁹ Pueden verse la Resolución de 21 de septiembre de 2016 de la Comisión de Garantía del Derecho de Acceso a la Información Pública (GAIP), que estima las Reclamaciones 123/2016 y 124/2016 (acumuladas) sobre el algoritmo matemático que determina la selección de los miembros de los tribunales correctores de las pruebas de acceso a la Universidad (PAU) y DNI de los candidatos presentados, y su Resolución 200/2017, de 21 de junio relacionada con las anteriores. Por lo que se refiere al CTBG, las resoluciones R/0058/2021 de 20 de mayo, sobre el acceso al algoritmo para calcular pensiones de la TGSS; RT/253/2021 de 19 de noviembre y RT/ 748/2021 de 10 de enero 2022 sobre acceso al código fuente de aplicación utilizada para el sorteo de tribunales asociados a procesos selectivos para educación en Madrid, y la Resolución del Expediente 551-2023 sobre el acceso a información del Sistema de Seguimiento Integral de los casos de Violencia de Género (Sistema VioGén).

⁵⁸⁰ Ello teniendo en cuenta que, como recuerda la sentencia de 30 de diciembre de 2021 del Juzgado Central de lo Contencioso-Administrativo núm. 8 (en el caso Bosco), «no existe ninguna norma que imponga a la Administración el desarrollo de aplicaciones con fuentes abiertas ni la adquisición de *software* libre».

carácter previo que está interactuando o siendo objeto de decisiones de una IA y no de una persona⁵⁸¹.

Según afirmaba el GT 29 en sus *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679*, sería suficiente con compartir *una explicación sobre el funcionamiento del sistema en lenguaje común para que cualquier persona interesada, sin los conocimientos técnicos suficientes, pudiera entender bajo qué parámetros se ha adoptado la decisión que le afecta, sin necesidad de revelar el código fuente o el secreto empresarial que lo protege*.

El mismo criterio se recogía en la *Carta ética europea sobre el uso de la IA en los sistemas judiciales* adoptada por la Comisión Europea para la Eficacia de la Justicia (en adelante, CEPEJ) en diciembre de 2018⁵⁸². En ella, partiendo de que la completa transparencia (por ejemplo, datos y código fuente abiertos) no será siempre posible debido al secreto empresarial, se proponen alternativas en pro de la transparencia algorítmica como la *explicación del sistema en un lenguaje claro y familiar mediante la descripción de la naturaleza de los servicios ofrecidos, las herramientas desarrolladas, su rendimiento y los riesgos de error, o la implantación de un sistema de auditoría y certificación de las metodologías utilizadas y su fiabilidad por parte de agencias o expertos independientes*⁵⁸³. Tales mecanismos de certificación y la posibilidad de realización de auditorías serían especialmente adecuados para controlar los resultados del algoritmo y comprobar si se están vulnerando derechos (no discriminación, protección de datos...), *sin necesidad de revelar la propia fórmula*.

En esa línea se sitúa también el Informe *Algorithms and Human Rights*, elaborado por el *Comité de Expertos en intermediarios de internet del Consejo de Europa (MSI-NET)*⁵⁸⁴, donde se reconoce que, si bien es improbable que en este contexto se ponga a disposición del público el código fuente de los algoritmos, sí podría requerirse que ciertas informaciones esenciales sobre los algoritmos fueran accesibles, por ejemplo, las variables utilizadas, los objetivos para los que se optimiza el diseño de los algoritmos, los datos de entrenamiento y los

⁵⁸¹ En este sentido se pronunciaba ya en el año 2019, NAVAS NAVARRO, S., «Herramientas LawTech, faltas de conformidad y fallos del algoritmo», *Diario La Ley*, N° 33, Sección Legal Management, 29 de Octubre de 2019, p. 4/10. Disponible en Laleydigital, LA LEY 13181/2019.

⁵⁸² La CEPEJ adoptó en diciembre de 2018, la Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno. Es la primera Carta adoptada por una organización europea sobre este tema y además va acompañada de un estudio científico, un glosario y una revisión de diferentes aplicaciones de inteligencia artificial que pueden contribuir a mejorar el trabajo de los profesionales jurídicos. Los cinco principios establecidos por la Carta son: 1. principio de respeto a los derechos fundamentales; 2. principio de no discriminación; 3. principio de calidad y seguridad; 4. principio de transparencia, imparcialidad y equidad; y 5. principio «bajo el control del usuario».

⁵⁸³ P. 9.

⁵⁸⁴ Accesible en <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5>

valores medios y desviaciones estándar de sus resultados o la cantidad y tipo de datos procesados por el algoritmo.

De hecho, se ha mantenido la posibilidad de utilizar ciertas técnicas matemáticas, como por ejemplo las copias, para averiguar las premisas que configuran el modelo sin que sea necesario acceder al código fuente ni a los datos de entrenamiento del algoritmo para analizar su potencial impacto discriminatorio⁵⁸⁵. Se afirma también que actualmente existen técnicas computacionales, como los compromisos criptográficos y las pruebas de conocimiento cero, que permiten verificar la regularidad o coherencia procedimental y proveer una explicación de lo que hace el sistema (mostrando cómo cada resultado o *output* es el resultado de aplicar la misma política o regla de decisión generada por el sistema a cada caso o *input* particular) sin necesidad de conocer cómo funciona su caja negra⁵⁸⁶.

Un mecanismo que se ha destacado por su idoneidad para garantizar la transparencia y salvar las objeciones de secreto empresarial o propiedad intelectual han sido las auditorías independientes⁵⁸⁷. Tal como mantiene RIVAS VALLEJO, «existe cierto consenso en la idea de que cabe mantener el secreto industrial del algoritmo mediante la auditoría de inteligencia artificial, introduciendo los mecanismos correctores oportunos si procede, pero salvando el código fuente del algoritmo al conocimiento público, pues existen métodos en la ciencia de la computación que permiten ejercer una supervisión sin comprometer dicho contenido si está sujeto a compromisos de propiedad o incluso de seguridad»⁵⁸⁸.

Otro sector doctrinal mantiene una postura más extrema y aboga por la prohibición de algoritmos de caja negra para que todas las decisiones sean explicables y revisables, opinión que a nuestro modo de ver no es en absoluto descabellada para los casos en que las decisiones de la IA tengan efectos directos en la vida de las personas. Es decir, cuando las decisiones algorítmicas (automatizadas o semiautomatizadas) puedan afectar a derechos fundamentales, es necesario que el diseño, desarrollo e implementación del sistema sea explicable, esto es, que se base en algoritmos de caja blanca que posibiliten la comprobación de posibles fallos o sesgos en su configuración o funcionamiento.

⁵⁸⁵ GINÈS i FABRELLAS, A., *op.cit.*, p. 315 y doctrina citada.

⁵⁸⁶ SOLAR CAYÓN, J.I., *op.cit.*, p. 154.

⁵⁸⁷ Así lo afirmaba el CTBG en su resolución R/0058/2021, ya comentada: «[m]ientras no se instauren otros mecanismos que permitan alcanzar los fines señalados con garantías equivalentes –como podrían ser, por ejemplo, *auditorías independientes u órganos de supervisión*–, el único recurso eficaz a tales efectos es el acceso al algoritmo propiamente dicho, a su código, para su fiscalización, tanto por quienes se puedan sentir perjudicados por sus resultados como por la ciudadanía en general, en aras de la observancia de principios éticos y de justicia».

⁵⁸⁸ RIVAS VALLEJO, P., «Herramientas desde el derecho antidiscriminatorio...», *cit.*, p. 555.

Y ello, aun a costa de sacrificar por el momento el uso de herramientas IA de mayor eficacia y precisión en estos casos⁵⁸⁹.

Por último, también se ha mantenido la necesidad de adoptar soluciones desde el punto de vista tecnológico para conseguir mayor transparencia algorítmica, que serán analizadas en un apartado *ad hoc*.

2.6. CONSIDERACIONES FINALES SOBRE LA NECESIDAD DE TRANSPARENCIA Y EXPLICABILIDAD PARA DETECTAR LA DISCRIMINACIÓN ALGORÍTMICA

Como se ha ido exponiendo, la transparencia y la explicabilidad algorítmicas cumplen diversas funciones: por un lado, ayudan a entender su funcionamiento, tanto para diseñadores y desarrolladores como para los afectados por sus efectos; por otro lado, permiten impugnar adecuadamente las decisiones construyendo argumentos sólidos para desvelar su posible ilegalidad; y por último, redundan en la confiabilidad y auditabilidad de los sistemas.

La transparencia ha sido una cuestión prioritaria a la que se ha dedicado especial atención por las instituciones y la doctrina desde los inicios del tratamiento regulatorio de la IA. De hecho, se ha erigido en un principio básico de muchos marcos normativos, éticos y de seguridad vigentes y proyectados. Es lógico, por ello, que también sea uno de los pilares básicos sobre los que se asienta el reciente Reglamento europeo que regula la materia y con el que se pretende fomentar la confianza en los sistemas inteligentes y mitigar riesgos como el sesgo y la discriminación injusta.

Si bien es cierto que la transparencia no puede impedir que los sistemas de IA lleguen a resultados errados y/o discriminatorios, ni garantizar su equidad y justicia, es fundamental para descubrir desviaciones que pasarían desa-

⁵⁸⁹ CHEONG, M., y otros, *op.cit.*, p. 14. En el Informe del World Economic Forum, *How to Prevent Discriminatory Outcomes*, p. 10, se pone un ejemplo al respecto. En ocasiones, no es necesario saber por qué un sistema de IA ha adoptado una decisión en un sentido u otro. Sencillamente, la transparencia no importa. Puede ser el caso de Google Maps donde no será preciso entender por qué se nos sugiere una ruta concreta. Sin embargo, cuando se trata de decisiones que impactan en derechos de las personas, la transparencia es imperativa. Por ejemplo, cuando un sistema de IA es el que decide sobre la concesión o no de la libertad condicional a un preso. En tales casos, identificar y remediar la posible discriminación dependerá de que se puedan conocer y comprender los criterios adoptados para llegar a una decisión concreta, es decir, debe poder trazarse paso a paso el camino que ha llevado al algoritmo a adoptar la decisión, las operaciones técnicas que realiza de inicio a fin del proceso. Por tanto, en ámbitos como la justicia penal, la vivienda, el bienestar social y la salud, no deberían desarrollarse o utilizarse sistemas de IA inexplicables o de caja negra.

percibidas si no se exigiera tanto en el diseño como en el despliegue y uso de los sistemas de IA⁵⁹⁰. Dicha exigencia debe operar en todo el ciclo de vida de los sistemas y ha de ser tanto interna como externa.

Es más, no solo ha de abogarse porque los sistemas sean transparentes, sino también, y sobre todo, porque sus resultados sean explicables. Mientras la transparencia tiene que ver con el proceso seguido para llegar a un resultado, y pretende dar a conocer y entender cómo se procesa el dato que se introduce en el sistema, la explicabilidad tiene que ver con el resultado de ese proceso y descende más al detalle⁵⁹¹. Por tanto, de lo que se trataría con la explicabilidad sería de que el afectado por la decisión pudiera conocer los motivos por los que se ha tomado una decisión que afecta a sus derechos y libertades y tuviera la posibilidad de impugnarlas. No sería suficiente con dar una explicación genérica sobre cómo funciona el proceso de toma de decisión, sino que habría que indicar los datos específicos de la persona en base a los cuales se ha tomado una decisión o se ha hecho un perfilado. En definitiva, no solo habría que informar de forma clara y sencilla de la denominada «lógica computacional» del sistema, sino también de su «lógica decisional». Sólo así podrán salvaguardarse derechos fundamentales como la tutela judicial efectiva, el derecho a la defensa o la igualdad y no discriminación.

En este sentido, es elogiable que el RIA haya acabado acogiendo un *derecho a la explicación* que viene a completar tanto las previsiones sobre transparencia y comunicación de información entre los diversos sujetos de la cadena de valor (principalmente, entre proveedores y responsables del despliegue) que ya preveía su versión inicial, como el insuficiente y criticado alcance del art. 22 RGPD que suponía una importante traba en la lucha contra la opacidad de los sistemas de IA.

Asimismo, cuando las decisiones se basan en datos personales, el control sobre nuestra información personal ha de implicar la posibilidad de ejercer, como mínimo, una serie de derechos como el de ser informados sobre el razonamiento subyacente al procesamiento de datos realizado por los algoritmos, el derecho a oponerse a dicho procesamiento y el derecho a recurrirlo. Ello no sólo cuando se trate de decisiones plenamente automatizadas sino también semiautomatizadas.

En la era de los algoritmos, donde la IA impacta en casi todos los aspectos de la vida de las personas, cada vez son más numerosas las medidas que se es-

⁵⁹⁰ Mantienen los expertos que la transparencia sobre los datos utilizados y sobre los efectos generados el esencial para evaluar su potencial impacto discriminatorio. Por todos, O'NEIL, *op.cit.*

⁵⁹¹ Así lo afirma el Director del Centro Europeo para la Transparencia Algorítmica (ECAT), Alberto Pena Fernández, en Diario La Ley, 8 de marzo de 2024.

tán adoptando en pro de la claridad algorítmica a fin de averiguar *cómo funciona el algoritmo* y, sobre todo, *bajo qué parámetros se ha adoptado la decisión que nos afecta*, lo que contribuirá a detectar y conseguir probar más casos de discriminación algorítmica. De hecho, se ha mantenido con acierto a nuestro juicio, que esta necesidad de garantizar el acceso a la información algorítmica para detectar si los sistemas discriminan podría considerarse una *cuestión de orden público*, ya que, averiguar si la IA discrimina no solo interesa a los particulares afectados sino también a los Gobiernos⁵⁹².

En definitiva, tanto la transparencia como la explicabilidad de los sistemas de IA que toman decisiones basadas en datos personales y con efectos jurídicos en personas concretas son cruciales para que la ciudadanía confíe en ellos. Como se analizará en el siguiente capítulo, para lograr estos objetivos debe adoptarse un enfoque interdisciplinar combinando medidas jurídicas, éticas y técnicas y tratando de encontrar un equilibrio adecuado entre los derechos de los ciudadanos y los intereses empresariales.

3. SEGUNDO OBSTÁCULO: PROBAR LA DISCRIMINACIÓN Y SU CAUSA

3.1. LA CARGA DE LA PRUEBA DE LA DISCRIMINACIÓN Y EL ESTÁNDAR DE PRUEBA

3.1.1. Consideraciones preliminares

De entrada, hay que dejar sentadas tres ideas importantes sobre la prueba de la discriminación:

En primer lugar, que la carga de descubrir la discriminación, iniciar procedimientos y probar el hecho dañoso recae, en principio, en el perjudicado por la decisión, conducta o práctica discriminatoria, esto es, en la víctima individual de la discriminación (o cuando lo permita la legislación nacional, en las organizaciones con un interés legítimo)⁵⁹³. La legislación actual en materia

⁵⁹² Así, LÓPEZ-TARRUELLA MARTÍNEZ, A., *op.cit.*, p. 195.

⁵⁹³ Por lo que se refiere a las acciones colectivas, en nuestro país es la Ley 15/2022 integral de igualdad, la que en su art. 29 reconoce legitimación activa a ciertos colectivos para la defensa del derecho de igualdad.

En su estudio sobre la discriminación algorítmica en Europa, GERARDS, J./ XENIDIS, *op.cit.*, p. 75, mantienen que los ordenamientos que carecen de mecanismos para iniciar acciones colectivas pueden no ser capaces de abordar adecuadamente los problemas y desafíos inherentes a la toma de

de igualdad y no discriminación se ha diseñado para ser alegada por los sujetos afectados para que el ordenamiento jurídico pueda reaccionar *a posteriori* frente a las decisiones discriminatorias, lo que no deja de ser paradójico cuando en muchos casos, como suele ocurrir en el ámbito de la discriminación algorítmica, la víctima ni siquiera es consciente de que está siendo discriminada.

En segundo lugar, que dicha prueba es bastante compleja tanto en los casos de discriminación analógica como algorítmica, porque quien discrimina no suele hacerlo abiertamente o de forma claramente identificable en atención a la reprochabilidad de su conducta. Por ello, puede llegar a convertirse en una *probatio diabólica*.

Por último, que dicha prueba exige a la parte demandante que aporte una «referencia comparativa», es decir, una persona que se halle en circunstancias sustancialmente iguales que ella y cuya principal diferencia respecto a la otra persona sea un motivo protegido. Se trata de convencer al órgano jurisdiccional de que la única explicación razonable de la diferencia de trato es la característica protegida de la víctima⁵⁹⁴.

En el escenario algorítmico, ese elemento de comparación no será fácil de conseguir porque exigirá acceder tanto al contenido del algoritmo (que

decisiones algorítmica. A su juicio, que compartimos, dado que la discriminación algorítmica es sistémica por naturaleza, los problemas que plantea podrían abordarse mejor a nivel colectivo que de forma individual. Sobre las acciones colectivas para hacer frente a la discriminación algorítmica nos remitimos al apartado 3.1.1. del Capítulo V.

⁵⁹⁴ Ejemplo de lo afirmado podría ser el asunto *Susanna Brunnhofer/Bank der österreichischen Postsparkasse AG*. En este caso la reclamante alegó discriminación por razón de sexo por percibir una retribución inferior a la de un compañero situado en el mismo nivel retributivo. El TJUE señaló que correspondía a la reclamante demostrar, en primer lugar, la percepción de una retribución inferior a la de su compañero y, en segundo lugar, la realización de un trabajo de igual valor, lo cual bastaría para crear la *presunción de que la diferencia de trato solo podía explicarse por razón del sexo de la reclamante*, cuya refutación correspondería al empleador. Vid., *Manual de legislación europea contra la discriminación*, cit. p. 49 (la cursiva es nuestra). Esta referencia comparativa es necesaria tanto en los casos de discriminación directa como indirecta. Por tanto, «tanto en los presuntos casos de discriminación directa como indirecta, los tribunales deben comparar, por ejemplo, los hombres con las mujeres, las parejas homosexuales con las parejas heterosexuales, o las personas con discapacidades con las personas sin discapacidades. Sin embargo, para la determinación de la discriminación indirecta, es necesario probar la existencia de dos grupos: uno en una posición de ventaja y otro en una posición de desventaja por la medida impugnada. En general, el grupo en situación de desventaja no está formado únicamente por personas con características protegidas. Por ejemplo, los trabajadores a tiempo parcial a los que coloca en una situación de desventaja determinada disposición son mayoritariamente mujeres, pero también puede haber hombres afectados. Por otra parte, no todas las personas con una característica particular sufren una situación de desventaja. Por ejemplo, en una situación en la que se exija un perfecto dominio de una lengua para poder acceder a un puesto, los candidatos extranjeros se encontrarán en general en una situación de desventaja, pero es posible que algunos de estos candidatos cumplan este requisito». Así se explica en la p. 64 del manual citado.

como vimos podrá estar protegido por derechos de propiedad intelectual o secreto empresarial, e incluso ser inescrutable o ininteligible), como a datos personales e información de otras personas (protegidos por la normativa *ad hoc* en la materia)⁵⁹⁵.

Respecto al problema de los datos personales, la solución podría ser solicitar esos datos una vez hayan sido anonimizados. Pero es que incluso si se pudiese contar con dichos datos por no ser personales o estar anonimizados, podría resultar muy difícil encontrar un elemento comparador (otra persona examinada bajo los mismos parámetros) teniendo en cuenta que algunos sistemas automatizados están en constante evolución y puede ser difícil o imposible encontrar dos casos en los que el algoritmo haya funcionado de manera idéntica⁵⁹⁶. Para sortear este problema se ha venido abogando por la necesidad de que esta tecnología contara con un sistema que permitiese registrar información sobre su funcionamiento, esto es, datos relativos a su movimiento, actuación y decisión⁵⁹⁷, exigencia que ha sido finalmente recogida en el RIA cuyo art. 12 dispone que los sistemas de alto riesgo deberán contar con «archivos de registro» que permitan registrar eventos automáticamente durante todo su ciclo de vida⁵⁹⁸.

⁵⁹⁵ Por ejemplo, en caso de que a una persona se le denegara un trasplante en favor de otra en base a la elaboración de un perfil realizado a partir de una lógica discriminatoria, para demostrar la discriminación la víctima tendría que aportar indicios con los que probar que la única explicación razonable de la diferencia de trato era la característica protegida. Es decir, que ante un historial clínico similar y otros datos personales tomados en cuenta para la elaboración del perfil en que se basa la posterior decisión, se descartó al sujeto por ejemplo por su raza (criterio al que se le atribuyó, directa o indirectamente, un peso negativo). Para probar tales extremos, sería preciso acceder a información sobre el algoritmo (variables introducidas, peso de cada una, etc.) y a datos de otros sujetos que se introdujeron en el sistema para su aprendizaje.

⁵⁹⁶ Así lo sostiene SORIANO ARNANZ, A., «La aplicación del marco jurídico europeo...», *cit.*, p. 78.

⁵⁹⁷ Como apuntaba NAVAS NAVARRO, S., «La perspectiva de género ...», *cit.*, la existencia desde el diseño de un «*event data recorder*» es esencial para obtener una justificación, aunque sea aproximada, que explique cómo y porqué un determinado sistema de IA ha llegado a una solución. Por otro lado, la información que proporcione esta caja negra puede ser importante en casos, por ejemplo, de responsabilidad civil por los daños causados por el sistema, en aras de facilitar la prueba a la víctima, tal como mantuvo el Grupo de Expertos de la Comisión en Responsabilidad y Nuevas Tecnologías, en su *Informe NTF 2019*. En él se mantenía la necesidad de que las tecnologías digitales emergentes se comercializaran con un registro de actividad apropiado a sus características y que la falta de dicho sistema de registro, o de un sistema de acceso razonable a los datos registrados, debería dar lugar a una inversión de la carga de la prueba para no perjudicar a la víctima. Y para SORIANO ARNANZ, A., «La aplicación del marco jurídico europeo...», *cit.*, p. 81, en los sistemas que se van actualizando, es necesario que los operadores guarden una copia de cada versión del programa utilizado en los procesos de toma de decisiones.

⁵⁹⁸ En el apartado 4 se añade que: «En el caso de los sistemas de IA de alto riesgo que figuran en el punto 1, letra a), del anexo III, las capacidades de registro incluirán, como mínimo: a) un registro del período de cada uso del sistema (la fecha y la hora de inicio y la fecha y la hora de finalización

3.1.2. La prueba de la discriminación

Sin perjuicio de este requisito técnico que han de cumplir ciertos sistemas de IA, las dificultades en torno a la prueba de la discriminación se han tratado de compensar en general aligerando la carga de la prueba al demandante. Este principio está muy consolidado y las Directivas europeas de igualdad obligan a los Estados Miembros a *«garantizar que corresponda a la parte demandada demostrar que no ha habido vulneración del principio de igualdad de trato cuando una persona que se considere perjudicada por la no aplicación, en lo que a ella se refiere, de dicho principio alegue, ante un tribunal u otro órgano competente, hechos que permitan presumir la existencia de discriminación directa o indirecta»*⁵⁹⁹. Por tanto, cuando el reclamante pueda acreditar la existencia de unos hechos de los que pueda presumirse la existencia de discriminación, corresponderá al presunto autor probar su inexistencia. En tal caso, deberá probar que la diferencia de trato no se basaba en un motivo protegido o que, aun siendo así, existía justificación objetiva y razonable para ello. Si el presunto autor de la discriminación es incapaz de probar cualquiera de estos dos extremos, deberá responder de los daños que la discriminación ocasione⁶⁰⁰.

En España, es el art. 30 de la Ley 15/2022 integral de igualdad el que regula la carga de la prueba al disponer que *«cuando la parte actora o el interesado alegue discriminación y aporte indicios fundados sobre su existencia, corresponderá a la parte demandada o a quien se impute la situación discriminatoria la aportación de una justificación objetiva y razonable, suficientemente probada, de las medidas adoptadas y de su proporcionalidad»*. Su Disposición Final 2^a también modificó la redacción del art. 217.5 LEC quedando su tenor literal como sigue: *«En aquellos procesos en los que la parte actora alegue discriminación y aporte indicios fundados sobre su exis-*

de cada uso); b) la base de datos de referencia con la que el sistema ha cotejado los datos de entrada; c) los datos de entrada con los que la búsqueda ha arrojado una correspondencia; d) la identificación de las personas físicas implicadas en la verificación de los resultados que se mencionan en el artículo 14, apartado 5».

⁵⁹⁹ Se recoge esta obligación en el art. 8 de la Directiva sobre igualdad racial, en el art. 10 de la Directiva relativa a la igualdad de trato en el empleo, en el art. 19 de la Directiva sobre igualdad de trato entre hombres y mujeres (refundición) y en el art. 9 de la Directiva de igualdad de mujeres y hombres en el acceso a bienes y servicios. Se trata de una cláusula estándar en la normativa comunitaria antidiscriminatoria ya existente en la derogada Directiva 97/80/CE del Consejo, de 15 de diciembre de 1997, relativa a la carga de la prueba en los casos de discriminación por razón de sexo (vid., art.4).

⁶⁰⁰ *Manual de legislación europea contra la discriminación*, cit. p. 264. Como afirma CHOCRÓN GIRÁLDEZ, A.M., «La acción civil para la defensa de los derechos derivados de la igualdad de trato y la no discriminación», *Indret: Revista para el Análisis del Derecho*, 4, 2023, p. 105, el demandado debe convencer al juzgador que su actuación no tiene un móvil discriminatorio. Vid., sobre el tema la STJUE de 25 de abril de 2013, asunto C-81/12 Asociația Accept y Consiliul Național pentru Combaterea Discriminării (ECLI:EU:C:2013:275).

tencia, corresponderá a la parte demandada la aportación de una justificación objetiva y razonable, suficientemente probada, de las medidas adoptadas y de su proporcionalidad [...]».

Nótese que dicha modificación vino a eliminar la referencia concreta a la discriminación *por razón de sexo* que se contenía en la redacción anterior –la cual traía causa de la Disposición adicional 5ª de la LOI y se limitaba a reiterar lo dispuesto en su art. 13–, pasando a abarcar todas las causas de discriminación. No obstante, la norma fue de nuevo modificada por la Disposición Final 5ª de la *Ley 4/2023, de 28 de febrero, para la igualdad real y efectiva de las personas trans y para la garantía de los derechos de las personas LGTBI*, que volvió a limitar la aplicación de la regla especial sobre la carga probatoria a los casos de discriminación por *ciertas causas* como el «sexo, la orientación e identidad sexual, expresión de género o las características sexuales». De esta forma, el tenor literal del art. 217.5 LEC vigente dispone que «[e]n aquellos procesos en los que las alegaciones de la parte actora se fundamenten en actuaciones discriminatorias por razón del sexo, la orientación e identidad sexual, expresión de género o las características sexuales, y aporte indicios fundados sobre su existencia, corresponderá a la parte demandada la aportación de una justificación objetiva y razonable, suficientemente probada, de las medidas adoptadas y de su proporcionalidad». Dicho cambio obliga a acudir al art. 30.1 de la *Ley 15/2022* integral de igualdad, en cuanto ley especial, para que se aplique la misma distribución de la carga de la prueba con independencia de la causa (o causas) de discriminación sobre la que se sustente la demanda⁶⁰¹.

A pesar de que la *Ley 15/2022* no deroga los preceptos sobre la carga de la prueba recogidos en otras normas especiales (a saber, art. 13 LOI, art. 32 *Ley 62/2003* y art. 77 *TRLGDPD*)⁶⁰², podría entenderse que es esta norma la que se ha de aplicar en los procesos civiles sobre discriminación en virtud del criterio de *lex posterior*. Con ella se cubre la laguna que afectaba a un buen número de causas de discriminación (la mayoría), respecto a las que, a falta de norma especial sobre inversión de la carga de la prueba, debía aplicarse la regla general de distribución de los arts. 217.2 y 3 LEC, salvo que el tribunal considerara conveniente, atendiendo a las circunstancias del caso concreto,

⁶⁰¹ CHOCRÓN GIRÁLDEZ, A.M., *op.cit.*, p. 102.

⁶⁰² De hecho, la *Ley 15/2022* no contiene ninguna disposición derogatoria, ni siquiera la usual derogación genérica. Según LOUSADA AROCHENA, J.F., «*Ley 15/2022, de 12 julio, integral para la igualdad de trato y la no discriminación: incidencia en el Derecho del Trabajo*», accesible en <https://elderecho.com/ley-15-2022-igualdad-trato-no-discriminacion-derecho-del-trabajo>, esto se debe interpretar en consonancia con su art. 7.II inciso final, donde se establece que «la presente ley consagra los niveles mínimos de protección y no perjudica las disposiciones más favorables establecidas en otras normas, debiendo prevalecer el régimen jurídico que mejor garantice la no discriminación».

aplicar los principios de disponibilidad y facilidad probatoria del art. 217.7 LEC.

En base a lo dispuesto en el art. 30 Ley 15/2022, ante un posible caso de discriminación, lo realmente importante para conseguir trasladar la carga de la prueba es la aportación de elementos suficientes o fundados de que realmente la decisión, criterio o práctica adoptada por el presunto autor es discriminatoria. Es decir, no basta con que el perjudicado alegue que la decisión es discriminatoria, sino que debe acreditar la *existencia de indicios que generen una sospecha, apariencia o presunción a favor de su alegato*⁶⁰³. Estos hechos habrán de ser valorados por los órganos jurisdiccionales nacionales con arreglo a sus normas procesales.

En puridad, más que ante una inversión de la carga de la prueba estaríamos ante un *reparto de la carga de la prueba* en tanto en cuanto el perjudicado no queda relevado de la carga de probar, ya que tendrá que realizar una primera e ineludible tarea probatoria y no meramente alegatoria⁶⁰⁴. Como se afirma,

⁶⁰³ Sobre esta carga de la prueba existe una consolidada jurisprudencia que parte de la STC de 23 de noviembre de 1981 (RTC\1981\38) según la cual «cuando se alegue que una determinada medida encubre en realidad una conducta lesiva de derechos fundamentales, incumbe al autor de la medida probar que su actuación obedece a motivos razonables y ajenos a todo propósito atentatorio de un derecho fundamental». En esta misma sentencia se destaca que para que exista este desplazamiento del «onus probandi» no basta simplemente con que el actor tache la medida de discriminatoria, sino que, además, ha de acreditar la *existencia de indicios que generen una razonable sospecha, apariencia o presunción a favor de su alegato*. Sólo en estos casos, la parte demandada asume la carga de probar la existencia de causas suficientes y reales para calificar de objetiva su decisión. No se le impone, por tanto, la prueba diabólica de un hecho negativo (la no discriminación) sino la de razonabilidad y proporcionalidad de la medida adoptada y su carácter absolutamente ajeno a todo propósito atentatorio de derechos fundamentales (contrario a la igualdad). Pueden verse también, entre otras, las SSTC 55/1983, de 22 de junio de 1983; 104/1987, de 17 de junio de 1987; 166/1988, de 26 de septiembre de 1988; 114/1989, de 22 de junio de 1989; 135/1990, de 19 de julio 1990; 21/1992, de 14 de febrero 1992; 7/1993, de 18 de enero de 1993; 226/1993, de 20 septiembre de 1993; 293/1993 de 18 de octubre de 1993; 85/1995, de 6 de junio de 1995; 136/1996, de 23 de julio; 87/1998, de 21 de abril; 29/2000, de 31 de enero; 114/2002, de 20 de mayo; 98/2003, de 2 de junio; 31/2014, de 24 de febrero.

⁶⁰⁴ En este sentido, CHOCRÓN GIRÁLDEZ, A.M., *op.cit.*, p. 104, quien considera que, más que una inversión de la carga probatoria, lo que establece realmente el art. 217.5 LEC es un «aligeramiento de la carga que, de ordinario, debería asumir el demandante». También RODRÍGUEZ ÁLVAREZ, A., «La carga de la prueba en la Ley integral para la igualdad de trato y la no discriminación: algunas consideraciones críticas desde la óptica del proceso civil», *Diario La Ley*, N° 10131, Sección Tribuna, 15 de Septiembre de 2022 (LA LEY 7764/2022), considera que el objetivo de estas disposiciones sólo es aligerar la carga de la prueba al actor, *pero en ningún caso exonerarle absolutamente de ella*. Para GARCÍA RUBIO, M.P., «La igualdad de trato entre hombres y mujeres y su repercusión en el derecho de contratos. Análisis del Proyecto de Ley Orgánica para la igualdad de mujeres y hombres a la luz de la Directiva 2004/113/CE», *Diario La Ley*, N° 6602, Sección Unión Europea, 30 de Noviembre de 2006, Año XXVII, Ref. D-256 (LA LEY 4083/2006), p. 6/21, no se trata de una inver-

por todas, en la STS 609/2015, 12 de noviembre de 2015: «*El demandante en este tipo de litigios sobre vulneración de derechos fundamentales no queda liberado de realizar toda actividad probatoria. Por el contrario, ha de probar lo que ha sido calificado como “indicios racionales” o “principio de prueba”. Una vez que el demandante ha realizado esa actividad, el demandado (empresario, Administración, etc.) deberá destruir la presunción, bien realizando “contraprueba” que destruya la base fáctica de la demanda, bien justificando que existe una causa justificadora de su actuación ajena a cualquier vulneración de derechos fundamentales*». En suma, ambas partes asumen cargas probatorias.

También se viene sosteniendo que, más que de aportar indicios, se trataría de presentar un *principio de prueba* o prueba *prima facie*, esto es, de la aportación de una serie de elementos que, aunque no fueran suficientes para dar por probada la discriminación, sirvieran para generar una apariencia favorable en cuanto a su veracidad⁶⁰⁵. Así se desprende de la jurisprudencia constitu-

sión de la carga de la prueba, sino de una regla de prueba conocida como prueba por presunciones: «Las presunciones no suponen, en principio, inversión probatoria, pues el actor ha de demostrar los datos (hecho base) en que se apoya la presunción (hecho deducido); no obstante, con ello la prueba de la discriminación se ve facilitada, con lo que, en la práctica, equivale a una inversión del *onus probandi*». También el TS en materia de responsabilidad civil ha dicho que invertir la carga de la prueba conlleva los mismos efectos que admitir la prueba de presunciones (*vid.* STS de 8 de julio de 1988). Respecto a las presunciones de hecho contempladas en el art. 386 LEC, explica EVANGELIO LLORCA, R., *op.cit.*, p. 573, que surgen dudas sobre si implican o no una inversión de la carga de la prueba. En opinión de la autora, que compartimos, la respuesta debe ser negativa habida cuenta de que el demandante sigue soportando la carga de probar los indicios en base a los cuales se trata de convencer al juez del hecho presunto (en nuestro caso, la posible existencia de discriminación), mientras que al demandado le corresponde probar (conforme a las reglas del art. 217 LEC) aquellos hechos o circunstancias que puedan desvirtuarlo para convencer al juez de que no existe un enlace preciso y directo, según las reglas del criterio humano, entre el hecho demostrado y el que es objeto de la presunción. Sigue diciendo la autora que, para LUNA YERGA, Á., *La prueba de la responsabilidad civil médico-sanitaria*, Madrid, 2004, pp. 307 y 308, las presunciones no llevan consigo una inversión de la carga de la prueba sino más bien una exoneración probatoria que, sin modificar las reglas generales de la carga de la prueba, sí favorece a la parte que tiene asignada en un primer momento la carga de la prueba del hecho del que se ve relevada. Y sobre la base de este argumento, añade que lo que esto significa, en su opinión, es que «no hay imposición de la carga de la prueba al demandado *ex ante*, sino que, siendo el actor quien tiene esa carga, si durante el procedimiento el demandante acredita un hecho base a partir del cual se pretende demostrar otro, el demandado tendrá que desvirtuar el hecho base o la presunción que de él deriva para evitar la condena. Sería algo así como una inversión de la carga de la prueba sobrevenida».

⁶⁰⁵ Así lo estima RODRÍGUEZ ÁLVAREZ, A., *op.cit.*, para quien «lo que se vendría a exigir sería la acreditación de un *fumus boni iuris* análogo al que se requiere para la adopción de medidas cautelares. Esta apariencia de buen derecho no prejuzgaría el fondo del asunto. Simplemente provocaría que la balanza que distribuye las cargas probatorias se decantase hacia el lado de la demandada. La aportación de este *fumus boni iuris* constituiría, por tanto, un *prius* lógico, un requisito previo para el traslado de la carga a la demandada». Mantiene esta autora que, aunque en la Ley 15/2022 se haga referencia a la necesidad de aportar indicios, en sentido estricto no se trata de «indicios» sino de un

cional, según la cual, *los hechos presentados deben ser de entidad suficiente como para considerar razonable la hipótesis de la vulneración de un derecho fundamental y conseguir influir en el ánimo del juzgador la sospecha fundada de un trato discriminatorio, sin obligar al demandante a acreditar el hecho base como si exige el indicio*⁶⁰⁶. En base

principio de prueba. Esto es así porque el indicio es el elemento que conforma el hecho base de toda presunción y una vez probado queda acreditado el hecho base que permite, a través del nexo lógico, llegar al hecho presumido. Aquí no nos hallamos ante una prueba por presunciones, ya que, si así fuera, la prueba habría de ser plena y, por ende, ni se aligeraría la carga de la actora, ni se produciría una situación de incertidumbre que requiriese la aplicación de la regla de juicio y, por extensión, de las normas que distribuyen la carga de la prueba. En similar sentido, CHOCRÓN GIRÁLDEZ, A.M., *op.cit.*, p. 104, afirma que para aliviar la carga de la prueba a la víctima se ha considerado que es más favorable que el actor presente una *evidencia inicial, un principio de prueba o prueba prima facie* en lugar de tener que aportar necesariamente una base indiciaria. Debe tenerse en cuenta que la prueba del indicio ha de ser directa, sin que pueda probarse a través de una presunción.

⁶⁰⁶ En la STC 49/2003, de 17 de marzo, se dice que «son admisibles diversos resultados de intensidad en la aportación de la prueba que concierne a la parte actora, pero deberá superarse inexcusablemente *el umbral mínimo de aquella conexión necesaria*, pues de otro modo, si se funda la demanda en alegaciones meramente retóricas o falta la acreditación de elementos cardinales para que la conexión misma pueda distinguirse, haciendo verosímil la inferencia, no se podrá pretender el desplazamiento del *onus probandi* al demandado». En la misma línea, entre otras muchas, la STC144/2006, de 8 de mayo, establece que «[p]ara apreciar la concurrencia del indicio tendrán aptitud probatoria, *tanto los hechos que sean claramente indicativos de la probabilidad de la lesión del derecho sustantivo, como aquéllos que, pese a no generar una conexión tan patente, y resultar por tanto más fácilmente neutralizables, sean sin embargo de entidad suficiente para abrir razonablemente la hipótesis de la vulneración del derecho fundamental*. Esto es, son admisibles diversos resultados de intensidad en la aportación de la prueba que concierne a la parte actora, pero habrá de superarse inexcusablemente el umbral mínimo de aquella conexión necesaria, pues de otro modo, si se funda la demanda en alegaciones meramente retóricas, o falta la acreditación de elementos cardinales para que la conexión misma pueda distinguirse, haciendo inverosímil la inferencia, no se podrá pretender el desplazamiento del *onus probandi* al demandado. En definitiva, el o la demandante que invoca la aplicación de la regla de la prueba indiciaria debe desarrollar una *actividad alegatoria suficientemente precisa y concreta en torno a los indicios de la existencia de discriminación*. Alcanzado, en su caso, el anterior resultado probatorio, sobre la parte demandada recaerá la carga de probar la existencia de causas suficientes, reales y serias, para calificar de razonable y ajena a todo propósito lesivo del derecho fundamental la decisión o práctica empresarial cuestionada, único medio de destruir la apariencia lesiva creada por tales indicios (STC 2/2009, de 12 de enero, FJ 3)».

En relación con la necesidad de acreditar esos indicios de discriminación, el TC ha sentado la siguiente doctrina: «Es sabido que la prueba indiciaria se articula en un doble plano (por todas, SSTC 90/1997, cde 6 de mayo y 66/2002, 21 de marzo). El primero, la necesidad por parte del trabajador de aportar un indicio razonable de que el acto empresarial lesiona su derecho fundamental, *principio de prueba o prueba verosímil dirigidos a poner de manifiesto el motivo oculto que se denuncia* (STC 207/2001, de 22 de octubre). Bajo esas circunstancias, *el indicio no consiste en la mera alegación de la vulneración constitucional, sino que debe permitir deducir la posibilidad de que ha podido producirse* (SSTC 87/1998, de 21 de abril; 293/1993, de 18 de octubre, 140/1999, de 22 de julio; 29/2000, de 31 de enero; 207/2001, de 22 de octubre; 214/2001, de 29 de octubre; 14/2002 de 28 de enero; 29/2002, de 11 de febrero y 30/2002, de 11 de febrero). *Sólo una vez cumplido este primer e inexcusable deber, recaerá sobre la parte demandada la carga de probar que su actuación tuvo causas reales absolutamente extrañas a la pretendida vulneración, así como que tenían entidad suficiente para justificar la decisión adoptada*. En otro caso, la ausencia de prueba empresarial

a esta doctrina se ha afirmado que «se atribuye al indicio un valor superior a la mera sospecha, pero inferior a la prueba»⁶⁰⁷. En consecuencia, parece que el estándar de prueba exigible para trasladar la carga de la prueba al demandado se ve considerablemente rebajado en el ámbito de la discriminación, ya que basta con que se dé una mera sospecha de su existencia⁶⁰⁸.

Cabría preguntarse si en los casos de discriminación múltiple e interseccional (que, como se dijo, son frecuentes en los casos de discriminación algorítmica) ese principio de prueba debería recaer sobre todas las causas de discriminación alegadas o bastaría con que se refiriera a una de ellas. En opinión de RODRÍGUEZ ÁLVAREZ, teniendo en cuenta que el art. 6.3 c) de la Ley 15/2022 establece que en tales supuestos «[...] la motivación de la diferencia de trato, en los términos del apartado segundo del artículo 4, debe darse en relación con cada uno de los motivos de discriminación», dicho principio de prueba debería recaer sobre todas y cada una de las causas de discriminación aducidas. Traslada la carga de la prueba al demandado, éste debería justificar su actuación con respecto a cada una de ellas y si consiguiera justificar su actuación respecto a uno de los motivos de discriminación, pero no en relación al otro u otros alegados, debería responder por la discriminación⁶⁰⁹.

A los efectos de poder valorar adecuadamente si hubo o no discriminación en el caso concreto en base a las pruebas presentadas por las partes, el segundo apartado del art. 217. 5 LEC dispone que «el órgano judicial, de oficio o a instancia de parte, podrá recabar informes o dictámenes de organismos públicos competentes». Esta medida viene a completar la regla especial so-

trasciende el ámbito puramente procesal y determina, en última instancia, que los indicios aportados por el demandante despliegue en toda su operatividad para declarar la lesión del derecho fundamental del trabajador» (STC 17/2003, de 30 de enero de 2003). En el mismo sentido puede citarse la STC 140/2014, de 11 de septiembre de 2014, donde añade el Tribunal que la finalidad de la prueba indiciaria en estos casos es evitar que la imposibilidad de revelar los verdaderos motivos del acto empresarial impida declarar que resulta lesivo del derecho fundamental y puntualiza que ese indicio razonable a aportar por el demandante es un *principio de prueba dirigido a poner de manifiesto el motivo oculto de aquél*.

⁶⁰⁷ Así, CHOCHRÓN GIRÁLDEZ, A.M., *op.cit.*, p. 104. Según señala el TS, «por definición, las pruebas indiciarias son aquellas que, sin ser inicialmente prueba propiamente dicha, sirven para fijar como ciertas unas presunciones alegadas para lograr el convencimiento de los hechos que se quieren demostrar» (STS 264/2024, de 19 de febrero de 2024)

⁶⁰⁸ Como explica PELÁEZ VARGAS, G., «Indicios y presunciones», *Revista Facultad de Derecho y Ciencias Políticas*, n. 48, 1974, p. 56, debido a que el indicio y la sospecha sirven para una inferencia, es frecuente que se otorgue a la sospecha la calidad de indicio y se utilice como medio de prueba. Afirma que la «sospecha es un juicio ligero, una inferencia que abre camino a la duda, como basada en un indicio en estado bruto, en un indicio (lato sensu) que requiere verificarse», y que, si bien el indicio y la sospecha, parten de un hecho indicador; «en el indicio tal hecho está plenamente probado, es una inferencia que tiene base o comprobación lógica y jurídica; en cambio la sospecha, carece de tal demostración, se funda en suposición, conjetura, apariencia».

⁶⁰⁹ RODRÍGUEZ ÁLVAREZ, A., *op.cit.*

bre la carga de la prueba en este ámbito y va dirigida a obtener información adicional o un mayor asesoramiento al juez a fin de poder dictar una resolución motivada⁶¹⁰. Por tanto, en el complejo ámbito de los sistemas de IA, se podrían solicitar informes y dictámenes de expertos en ciencias de la computación que ayudaran al tribunal a comprender la lógica decisional del sistema respecto del que se sospecha que arroja resultados discriminatorios.

3.1.3. Los indicios en el ámbito de la discriminación algorítmica

Sea como fuere, en el contexto de la discriminación algorítmica dichos *indicios o sospechas fundadas* de discriminación no serán fáciles de conseguir teniendo en cuenta la complejidad y opacidad de los sistemas, sobre todo si no se han realizado auditorías independientes o evaluaciones de impacto en los derechos fundamentales, o si las estadísticas sobre el impacto que tiene el algoritmo sobre determinados grupos no están disponibles o los operadores se niegan a revelar información sobre ellas o sobre el sistema en general⁶¹¹. Por eso, en este ámbito se hace especialmente necesario que la parte aparentemente responsable facilite cierto grado de transparencia sobre el sistema, ya que sin dicha cooperación será muy difícil reunir esos indicios razonables de que existe discriminación. En algunos casos, sólo se requerirá transparencia suficiente para acceder a los resultados generados para otras personas y conseguir un elemento de comparación; en otros casos, será necesaria una transparencia total del sistema para poder demostrar la discriminación y determinar si es directa o indirecta⁶¹².

Baste recordar aquí, que mientras la discriminación directa es aquella en que se da un *trato* peor a una persona respecto de otra que se encuentre en una situación comparable en base exclusivamente a una característica prote-

⁶¹⁰ Afirma CHOCRÓN GIRÁLDEZ, A.M., *op.cit.*, p. 106 que dichos organismos suelen ser entidades gubernamentales con experiencia en la promoción y protección de derechos fundamentales o expertos técnicos en áreas de conocimiento específicas, y que, en la medida en que tales dictámenes son un medio de aportar máximas de experiencia que aumenten la capacidad de juicio del órgano judicial, guardan estrecha vinculación con la prueba pericial.

⁶¹¹ GERARDS, J./XENIDIS, R., *op.cit.*, p. 144. En p. 119 se apunta que las características particulares de los algoritmos conllevan que resulte muy gravoso para las víctimas individuales (e incluso para los órganos de control) demostrar que el sistema perjudica a un grupo protegido y establecer un caso *prima facie* de discriminación. Es decir, la oscuridad de los procesos algorítmicos de toma de decisiones dificulta la obtención de la información necesaria para mostrar una desventaja significativa para un grupo protegido.

⁶¹² En cualquier caso, habría que sortear los problemas sobre protección de datos personales y secreto empresarial o propiedad intelectual, anteriormente mencionados, amén de los inherentes al propio aprendizaje autónomo del algoritmo.

gida, la indirecta se produce cuando se adopta una medida, criterio o disposición, aparentemente neutras, pero que genera un *impacto* negativo sobre un colectivo protegido, sin que exista justificación objetiva para ello⁶¹³.

Siendo esto así, la prueba de la discriminación algorítmica *directa* exigiría demostrar que el algoritmo decidió en base a una característica protegida de la persona, lo que podría ocurrir bien porque a la hora de diseñarlo se introdujo como variable a tener en cuenta dicha característica a la que se otorgó un peso negativo, bien porque a pesar de no haberse introducido, el algoritmo la hubiera inferido a partir de otros datos no relevantes para predecir la variable objetivo adjudicándole también un valor negativo.

Como explica SORIANO ARNANZ, en los casos en que el algoritmo tuviera en cuenta explícitamente la pertenencia al grupo protegido, podría ser suficiente con que una persona interesada solicitase una explicación acerca de la lógica subyacente al sistema para que se pudiera detectar la existencia del trato discriminatorio; sin embargo, en los casos en que es el propio algoritmo el que infiere la pertenencia al grupo protegido a partir de otros puntos de datos, será difícil probar que se trata de una discriminación directa si no se tiene acceso al sistema⁶¹⁴.

Además, considera esta autora que los casos de discriminación directa más sencillos de probar, incluso sin acceder al código fuente, serían aquellos en que todos los miembros del grupo desaventajado recibieran un trato peor en comparación con aquellas personas no pertenecientes al grupo. Es decir, si estando en una situación análoga, el algoritmo sistemáticamente arrojara resultados peores para las personas pertenecientes a un grupo protegido que para aquéllas que no pertenecen a él, sería más sencillo detectar y probar la discriminación directa⁶¹⁵. Por el contrario, cuando la decisión discriminatoria no afectase a todas las personas de un grupo sino a una o varias de ellas, sin

⁶¹³ Todas las Directivas europeas de igualdad definen de forma similar o incluso idéntica ambas modalidades discriminatorias variando tan solo la razón por la que se discrimina (vid., art. 2 Directiva sobre la igualdad racial; art. 2.2.a. Directiva sobre igualdad en el empleo; o art.2.1.a. Directiva sobre igualdad de trato entre hombres y mujeres).

⁶¹⁴ SORIANO ARNANZ, A., «La aplicación del marco jurídico europeo...», *cit.*, p. 74. Advierte la autora que, en ocasiones, aunque se tenga acceso al sistema tampoco se podrá probar que la discriminación es directa. Esto podría ocurrir cuando el sistema fuera tan incomprensible que no se pudieran identificar con claridad las variables empleadas, ni si se ha tenido en cuenta de manera explícita o no la característica protegida. La consecuencia podría ser que casos que deberían ser considerados como de discriminación directa, acabarían calificándose como indirecta por imposibilidad de probar la primera.

⁶¹⁵ Sería el caso, por ejemplo, de que un sistema de *credit scoring* rechazara automáticamente a todos los candidatos de una raza determinada a pesar de tener un historial crediticio muy similar o prácticamente idéntico al de otras personas de otra raza a los que se les concedió el préstamo.

acceso al sistema sería más complicado probar que estamos ante un caso de discriminación directa porque se podría alegar que la diferencia de trato entre dos personas que se encontraban en una posición equivalente se basó en un motivo distinto a la pertenencia de una de ellas al grupo desaventajado. Es decir, se podría mantener que el algoritmo estaría tomando las decisiones basándose en criterios objetivos y no en características protegidas⁶¹⁶. Por tanto, sería muy difícil probar en estos casos que la causa del tratamiento menos favorable es la pertenencia de esa persona al grupo protegido⁶¹⁷. Esta situación podría darse cuando, a pesar de que el sistema atribuyera valor negativo a una característica protegida, el resultado o decisión final no fuese discriminatoria al haber tenido en cuenta otras muchas variables o características de la persona tras cuya ponderación se obtiene un resultado favorable⁶¹⁸.

Por lo que se refiere a la discriminación algorítmica *indirecta*, la disposición, práctica o criterio aparentemente neutro podría ser también una variable específica introducida en el sistema que, por el valor que éste le otorga, genere un resultado discriminatorio para el grupo desaventajado⁶¹⁹. Para

⁶¹⁶ Expone también este problema RIVAS VALLEJO, P. «Análisis...», *cit.*, pp. 426-427, en los siguientes términos: «cuando el sistema de decisión excluye a grupos de individuos distintos con características diversas protegidas (o incluso no protegidas), la comparación entre la persona excluida (la reclamante) y las elegidas o priorizadas no permitirá constatar que la causa de exclusión sea únicamente la que esta posee, pues confluirá todo un universo de características inferidas de los datos analizados igualmente rechazadas. Concurriendo por ejemplo el binomio hombres-mujeres como posible fuente de la discriminación, el análisis se simplifica, al poder centrarse en la exclusión sistemática o en un alto porcentaje de las mujeres, pero, si se trata de determinar la exclusión y su causa cuando concurren características correspondientes a distintos grupos protegidos, a una multitud de ellos (discapacidad, etnia, religión, procedencia, y otros posibles rasgos), pero no otros o no su combinación con otros, concluir la existencia de sesgo (e incluso de su indicio) se presenta como una tarea harto compleja, en tanto la multitud de variables permite considerar también que han primado otros rasgos a priori no detectados».

⁶¹⁷ SORIANO ARNANZ, A., «Discriminación algorítmica: garantías...», *cit.*, p. 153. En estos casos, sigue diciendo la autora, estimar o no la discriminación dependerá de lo evidente que sea el elemento discriminatorio en la decisión adoptada y de cómo lo interprete cada tribunal.

⁶¹⁸ SORIANO ARNANZ, A., «Discriminación algorítmica: garantías...», *cit.*, p. 152, pone como ejemplo el caso del algoritmo utilizado por el servicio de desempleo austriaco. Dicho sistema valora una pluralidad de categorías de datos entre las cuales ser mujer resta puntos, pero ser nacional austriaca o de otro Estado miembro suma puntos. Por tanto, es posible que a pesar de exista un trato discriminatorio en la toma de la decisión, el resultado final no sea discriminatorio si en una persona concurren pocos datos que restan puntos.

⁶¹⁹ Este supuesto podría acaecer, tal como señala RIVAS VALLEJO, P. «Análisis...», *cit.*, pp. 426-427, cuando el fundamento de la decisión del algoritmo se basa en variables como la *distancia del lugar de trabajo*. Se trataría de un criterio aparentemente neutro que no determinaría *per se* la existencia de discriminación. Ahora bien, si tal factor se vinculara a una procedencia concreta, y ésta a una característica protegida, la inferencia realizada por el algoritmo derivaría en una decisión discriminatoria (por ejemplo, los individuos que residen en la zona rechazada son mayoritariamente población inmigrante). No obstante, «tan complejo y largo proceso podría provocar la ruptura del

determinar la variable o variables que generan el resultado discriminatorio, también se necesitaría tener acceso al sistema. Y como ya dijimos, en ocasiones la gran cantidad de datos procesados y la complejidad estructural de estos sistemas de procesamiento impiden identificar de manera clara cuáles son los elementos que condicionan sus resultados.

Pues bien, si la opacidad técnica o jurídica del sistema impidiera determinar qué variable o variables o qué inferencias pueden haber ocasionado el resultado discriminatorio, la solución podría pasar por considerar al propio algoritmo en sí mismo (al sistema en su totalidad) como disposición, práctica o criterio aparentemente neutro, siempre y cuando se consiguiera demostrar su impacto discriminatorio, para lo cual serían especialmente importantes las pruebas estadísticas. De hecho, un sector doctrinal no alberga dudas de que los casos de discriminación algorítmica provocarán un aumento del empleo de este tipo de prueba⁶²⁰.

A la vista de las anteriores consideraciones, no cabe duda de que es necesario cierto grado de transparencia sobre el sistema para averiguar si el mismo discrimina directa o indirectamente.

nexo de conexión necesario para establecer la preceptiva relación entre la decisión y la discriminación pretendida, precisamente por la distancia entre los datos de referencia y la consecuencia analizada. *En definitiva, la propia dinámica de funcionamiento de los algoritmos introduce una dosis mucho mayor de dificultad en la detección de sesgos discriminatorios, que obliga a perfeccionar la granularidad del análisis*» (la cursiva es nuestra).

⁶²⁰ SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación...», *cit.*, p. 11 de 31; MUÑOZ GUTIÉRREZ, C., *op.cit.*, p. 289; EGUÍLUZ CASTAÑEIRA, J.U., *op.cit.*, p. 337; FERNÁNDEZ HERNÁNDEZ, C., *op.cit.*, p. 4/8. Conviene recordar la importancia que otorgan los tribunales a la evidencia estadística como indicio de discriminación indirecta para aplicar la inversión de la carga de la prueba: SSTJUE 28 de febrero de 2013 (caso c-427/11, asunto *Kenny et al.*); 3 de octubre de 2019 (asunto *Schuch-Ghannadan*); o 16 de julio de 2015 (asunto C-83/14, *Chez Razpredelenie Bulgaria AD*), donde se indica que es posible que la discriminación indirecta se establezca por cualquier medio, incluso basándose en pruebas estadísticas. Nuestro TC también ha destacado la importancia de las estadísticas en sentencias como la 91/2019, de 3 de julio (RTC 2019\91) donde se afirma que «la forma de abordar el análisis de la discriminación indirecta por razón de sexo pasa necesariamente por atender a los datos revelados por la estadística (SSTC 128/1987, FJ 6, y 253/2004, FJ 8)». En el caso concreto, «la consulta de estadísticas actuales en el mercado laboral indica que el porcentaje de mujeres que desempeñan un trabajo a tiempo parcial es claramente superior al de los hombres [...]» y «estos datos permiten confirmar, quince años después, la conclusión que ya extrajimos en la STC 253/2004, de que el contrato a tiempo parcial es una institución que afecta de hecho predominantemente al sexo femenino, lo que obliga, como señalamos entonces, a «examinar con mayor cautela el impacto de la regla sobre cómputo de periodos de carencia contenida en la norma cuestionada, pues, acreditada estadísticamente la realidad sociológica indicada, esto es, que la medida afecta a una proporción mucho mayor de mujeres que de hombres (*disparate effect*) para descartar la existencia de una discriminación indirecta prohibida por el art. 14 CE, habría que constatar que esa medida se explica por razones objetivamente justificadas de política social sin carácter discriminatorio».

En los casos en que el operador no diera acceso ni información sobre el sistema algorítmico (negativa que podría estar justificada por estar protegido por derechos de propiedad intelectual o como secreto empresarial, o ser injustificada⁶²¹), la doctrina propone que al menos debería darse la posibilidad de experimentar con el programa, de forma que pudieran introducirse diferentes perfiles para comparar los resultados entre unos y otros y poder así comprobar si discrimina o no con base en alguna categoría protegida⁶²². Es decir, debería de poder exigirse la realización de ciertas pruebas para descubrir cómo clasifica el algoritmo a partir de los resultados que proporciona. Por tanto, lo que habría que hacer sería probar qué resultados arroja el algoritmo ante datos de entrada ligeramente diferentes. Estos casos habría que diseñarlos en busca de esos diferentes resultados a partir de las sospechas que se tuvieran. Por ejemplo, si para evaluar la solvencia o el perfil de riesgo de los clientes de un banco y decidir si se les concede o no un préstamo se estuviera utilizando un algoritmo de aprendizaje supervisado no paramétrico que diera como salida un árbol de decisión, lo primero que habría que hacer ante la posible sospecha de que el algoritmo está discriminando, por ejemplo, a las mujeres de raza negra, sería analizar el resultado del árbol que genera ese algoritmo. Dicho resultado, al ser una salida y no el algoritmo en sí, no estaría protegido. Se podría solicitar la auditoría de casos concretos para comprobar si terminan en concesión o denegación sólo en base a las características o motivos protegidos. No obstante, la cuestión no sería tan sencilla en tanto en cuanto esas salidas (reglas o árboles generados) podrían ir cambiando en función de los datos de entrada suministrados⁶²³.

⁶²¹ Pensemos que, según mantiene la doctrina que ha estudiado el tema, actualmente hay técnicas matemáticas que permiten dar información sobre las premisas que configuran el modelo sin acceder a su código fuente ni a los datos de entrenamiento para analizar su impacto discriminatorio, como por ejemplo las copias. Así lo afirma GINÈS i FABRELLAS, A., *op.cit.*, p. 315 y doctrina citada.

⁶²² Esta posibilidad la propone, entre otros, SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación...», *cit.*, p. 10/31. También sostiene CASTELLANOS CLARAMUNT, J., «Derecho e inteligencia artificial: atención especial a los sesgos, la privacidad y la protección de datos», 2023, que entre las estrategias para detectar sesgos en la IA, los expertos mencionan la realización de pruebas o experimentos que permitan evaluar el comportamiento del sistema en diferentes situaciones y contextos, y verificar si existen diferencias en su desempeño en función de variables como el género, la raza, la edad o la orientación sexual de las personas que interactúan con él. Recuperado de <https://idpbarcelona.net/derecho-e-inteligencia-artificial-atencion-especial-a-los-sesgos-la-privacidad-y-la-proteccion-de-datos> (última consulta 3 de junio de 2024)

⁶²³ Sin perjuicio de que esto pueda servir para obtener ese principio de prueba, indicio o sospecha necesaria para trasladar la carga de la prueba al presunto autor de la discriminación, debe advertirse que para identificar los sesgos discriminatorios en los algoritmos no basta con observar las decisiones finales, sino que sería necesario realizar un análisis más detallado de cada paso del proceso, desglosando los datos y evaluando cómo el algoritmo utiliza las variables. Esa granularidad del análisis permitiría identificar cómo el algoritmo pondera ciertas variables, evaluar si las correlaciones entre variables están

Si en todas las pruebas que se hicieran con mujeres, la concesión o denegación de los préstamos se repartiese de forma balanceada, si con las personas de raza negra que solicitaran un préstamo ocurriera lo mismo, pero entre las mujeres negras se diera casi un 100% de denegaciones, podríamos considerar que dicho resultado supondría un *indicio razonable* de que el algoritmo discrimina a dicho colectivo, siempre y cuando la muestra hubiese sido suficientemente representativa, ya que, en caso contrario, dicha denegación podría deberse a cualquier otro factor (por ejemplo, que eran morosas)⁶²⁴.

Surgen dudas sobre cómo medir la existencia de discriminación, es decir, cuáles serían los datos estadísticos suficientes para constituir indicios de discriminación y, por tanto, desplazar la carga de la prueba. Al analizar las pruebas estadísticas del efecto desproporcionadamente perjudicial para el «grupo protegido» respecto a otras personas en situación similar, el TJUE y el TEDH buscan muestras de que una proporción especialmente grande de los perjudicados pertenezca a dicho grupo, pero sin fijar un porcentaje concreto y analizando cada situación en particular⁶²⁵. La sentencia del TJUE de 9 de febrero de 1999, admitió incluso la posibilidad de apreciar la concurrencia de discriminación cuando «los datos estadísticos mostraran una diferencia menos importante, pero persistente y relativamente constante durante un largo período de tiempo», constancia y uniformidad en los resultados que son propias de los procesos automatizados de toma de decisiones⁶²⁶.

introduciendo sesgos o examinar cómo diferentes subgrupos (por ejemplo, mujeres de cierta etnia o personas mayores con cierto rango de ingresos) se ven afectados por sus resultados.

⁶²⁴ Para RIVAS VALLEJO, P. «Análisis...», *cit.*, p. 425, el sesgo del algoritmo, de existir, va a provocar un resultado fácilmente acreditable a través de la comparación entre los individuos elegidos y los excluidos, suficiente para generar la duda necesaria. Para GINÈS i FABRELLAS, A., *op.cit.*, p. 313, «constituye indicio de discriminación indirecta diferencias de sexo, origen racial u otras causas protegidas en el impacto del algoritmo. Así, por ejemplo, comparando el número de solicitudes recibidas con el número de personas admitidas en un proceso de selección en sus distintas fases, la existencia de una menor ratio para las mujeres en comparación con la de los hombres. Similarmente, también podría constituir indicio de discriminación algorítmica la existencia de diferencias sexuales, raciales u otras causas protegidas en el porcentaje de error o falsos positivos o negativos del algoritmo» Y según TOLOSA, P./DIBO, C., *op.cit.*, p. 161, cuando se trata de algoritmos de caja negra no se puede saber cómo se llegó a la decisión de rechazar sistemáticamente a una persona en un proceso de selección, pero para tratar de demostrar la discriminación por razón de género, se podría probar que no había ninguna mujer contratada para cumplir las funciones del puesto ofertado.

⁶²⁵ Sobre este tema, vid., *Manual de legislación europea contra la discriminación*, *cit.*, p. 275 y ss.

⁶²⁶ TJUE, C-167/97, *Regina/Secretary of State for Employment, ex parte: Seymour-Smith y Pérez*, 9 de febrero de 1999. No obstante, en relación con los hechos concretos de este caso, el TJUE señaló que las estadísticas presentadas, que indicaban que el 77,4 % de los hombres y el 68,9 % de las mujeres se ajustaban al criterio objeto de controversia, no acreditaban que un porcentaje considerablemente menor de mujeres pudiera cumplirlo.

En relación con la necesidad de aportar datos concretos o estadísticas para probar la discriminación, debe tenerse en cuenta lo que mantiene el TJUE en su sentencia de 3 de octubre de 2019, caso *Schuch-Ghannadan*, en relación con el art.19 de la Directiva 2006/54/CE, de 5 de julio de 2006⁶²⁷. Señala el tribunal que dicho precepto debe interpretarse en el sentido de que no exige que el perjudicado por la discriminación presente estadísticas o hechos concretos para demostrar una presunción de discriminación, si no tiene acceso a tales estadísticas o hechos o únicamente puede acceder a ellos con muchas dificultades. Será suficiente, en tales casos, con que *fundamente la presunción de discriminación basándose en datos estadísticos generales*.

Cuando se emplearan sistemas de IA, se podría proponer un procedimiento para medir o evaluar la probabilidad de que exista discriminación algorítmica a partir de la fijación de un *score* de riesgo. En dicho procedimiento habría que tener en cuenta, entre otros extremos, si el algoritmo empleado es de caja blanca o negra, qué variables se han introducido o cuál ha sido la muestra de entrenamiento y si esa selección ha sido aleatoria o no. Si al pasar todas las pruebas en *score* de riesgo se superara cierto umbral, habríamos conseguido crear esa sospecha de que el sistema discrimina, suficiente para trasladar la carga de la prueba a la otra parte.

En definitiva, si de las pruebas realizadas con el programa y estadísticas obtenidas puede extraerse que el sistema tiene un efecto o impacto desproporcionadamente perjudicial para el grupo protegido respecto a otras personas en situación similar, se podría considerar que existe un *caso prima facie de discriminación* con el que se conseguiría desplazar la carga de la prueba hacia el presunto autor de la discriminación⁶²⁸.

Lo relevante a estos efectos vendría a ser el resultado discriminatorio del sistema con independencia de la causa concreta que lo provocó. Es decir, no sería necesario probar, por ejemplo, cómo el algoritmo llegó a inferir o establecer una correlación entre una variable aparentemente inofensiva y una causa de discriminación prohibida, ya que esto sería muy difícil o incluso imposible en modelos de caja negra. No obstante, como ha apuntado algún au-

⁶²⁷ Tribunal de Justicia de la Unión Europea (Sala Tercera) de 3 de octubre de 2019 (asunto C-274/18). Vid., concretamente, la tercera cuestión prejudicial planteada.

⁶²⁸ Explica SÁEZ LARA, C., *op.cit.*, p. 50, que en EEUU se propone recurrir a la denominada doctrina de la discriminación *per se* de los sistemas algorítmicos con efecto adverso sobre las categorías protegidas, para conseguir trasladar al demandado la carga de probar la necesidad de tales prácticas con efectos discriminatorios: «Esta doctrina parte de constatar, de un lado, que la línea de la responsabilidad (también empresarial) por los problemas creados por los algoritmos de aprendizaje automático es borrosa y, de otro, que existen dificultades de prueba, para que pueda imputarse responsabilidad a la empresa, en caso de falta de intención de discriminar o de falta de certidumbre sobre los resultados sesgados de un algoritmo».

tor, aun cuando la mera existencia de un impacto discriminatorio constituiría el indicio necesario para desplazar la carga de la prueba, no sería irrelevante el proceso que conduce a ese resultado ni los factores que intervienen en él, puesto que ambos extremos servirían para determinar la gravedad del incumplimiento y el alcance de la responsabilidad⁶²⁹. Por ejemplo, sería una negligencia grave merecedora de una responsabilidad mayor diseñar algoritmos que tengan en cuenta, como variables para la toma de decisión, ciertos criterios protegidos a los que se otorga, además, un peso negativo⁶³⁰.

En los casos en que el operador se negase a proporcionar cualquier información sobre el algoritmo e incluso a dejar que se experimentase con el programa, amén de que dicha negativa sería *a priori* muy significativa, la misma podría tener importantes efectos:

En primer lugar, si estuviéramos ante un caso de decisiones automatizadas del art. 22 RGPD, la negativa del responsable a proporcionar información algorítmica supondría una vulneración de la normativa de protección de datos por cuanto, como ya se dijo, en ella se contempla un derecho de información reforzada a favor del interesado en virtud del cual el responsable debe explicar la lógica aplicada por el algoritmo (y a nuestro juicio, también la lógica decisional) en lenguaje común para que cualquier persona, sin los conocimientos técnicos suficientes, pudiera entender bajo qué parámetros ha podido verse afectado por un modelo matemático de decisión⁶³¹.

Además, también supondría una vulneración del nuevo derecho a una explicación consagrado en el art. 86 RIA, en virtud del cual, aquel que se vea afectado por una decisión que el responsable del despliegue adopte basándose en los resultados de un sistema de IA de alto riesgo y que produzca efectos jurídicos o le afecte considerablemente del mismo modo, tendrá derecho a obtener del responsable del despliegue explicaciones claras y significativas acerca del papel que el sistema de IA ha tenido en el proceso de toma de decisiones y los principales elementos de la decisión adoptada.

⁶²⁹ Así lo estima RIVAS VALLEJO, P. «Análisis...», *cit.*, p. 427, para quien, aunque en nuestro Derecho prima el resultado, la conducta en sí también es relevante.

⁶³⁰ En el estudio sobre discriminación algorítmica en Europa llevado a cabo por GERARDS, J./XENIDIS, R. ya citado, se pone de relieve que, por ejemplo en Alemania, la doctrina considera que para que la distribución de la responsabilidad sea justa, debería responder el presunto autor de la discriminación cuando hubiera indicios de que el error del programa era fácilmente detectable, es decir, que pudiendo evitarse la discriminación no se hizo. No obstante, consideran mucho más recomendable para evitar estos supuestos que se adopten soluciones técnicas de «antidiscriminación por diseño», ya que las respuestas legales *ex post* requerirían mayor tiempo para implementarse.

⁶³¹ Parte de la doctrina considera que esta información podría ser suficiente para fundamentar una eventual demanda por discriminación algorítmica.

En segundo lugar, podría constituir un indicio suficiente de que existe discriminación en virtud de la doctrina sentada por el TJUE en casos de discriminación tradicionales como el asunto C-415/10 (*Galina Meister/Speech Design Carrier Systems GmbH*) de 19 de abril de 2012 (en adelante, caso Meister). En dicho caso, la demandante alegó discriminación por razón de sexo, edad y origen étnico en un proceso de contratación y trató de obtener información sobre quiénes eran los otros solicitantes de empleo. La empresa se negó a proporcionar dicha información, impidiendo de facto que la solicitante estableciera un caso *prima facie* de discriminación. El TJUE determinó que el Derecho de la Unión no preveía ningún derecho de los trabajadores que reunieran las condiciones enunciadas en la oferta de trabajo y cuya candidatura no hubiera sido seleccionada, a recibir una justificación de los motivos por los que se hubiera seleccionado a un candidato diferente al término del proceso de selección. Sin embargo, por lo que a nosotros interesa, también indicó que *la negativa de la empresa a proporcionar dicha información a la solicitante, impidiéndole así reunir indicios suficientes para invertir la carga de la prueba, podía considerarse por los tribunales nacionales como uno de los factores que, junto con otros, permitieran presumir la existencia de discriminación en el proceso de selección*⁶³².

En el asunto 109/88 (*Handels-og Kontorfunktionærernes Forbund i Danmark/Dansk Arbejdsgiverforening*, en nombre de *Danfoss*) –en adelante, caso Danfoss–, el TJUE apreció que cuando una empresa aplica un sistema de remuneración que carece absolutamente de transparencia, corresponde al empresario demostrar que el mismo no es discriminatorio⁶³³. Esta doctrina de la falta de transparencia, sin obligar al empresario a divulgar el código fuente del algoritmo, le trasladaría la carga de probar que no existió trato discriminatorio, o que el mismo era justificable.

En el asunto C-274/18 (*Minoo Schuch-Ghannadan contra Medizinische Universität Wien*) –en adelante, caso Schuch-Ghannadan–, a fin de facilitar la carga de la prueba al demandante, considera que no se puede considerar carente de fundamento una presunción de discriminación por razón de sexo por el mero hecho de que la parte demandante, a falta de datos más precisos de los que no disponía o a los que le resultaba difícil acceder, se hubiera ba-

⁶³² *Manual de legislación europea contra la discriminación*, cit., p. 268. Opina ALAMEDA CASTILLO, M.T., *op.cit.*, p. 31, nota 65, que esta sentencia parece avanzar una suerte de potestad o facultad de información del candidato presuntamente discriminado, pero sin perfilar, sin embargo, su alcance y la potencial afectación de derechos de otros candidatos.

⁶³³ Si bien el Tribunal atribuye a la parte demandada la carga de probar que un sistema de remuneración opaco no es discriminatorio, solo lo hace una vez que se ha demostrado que dicho sistema perjudica a las mujeres, es decir, una vez que los trabajadores femeninos consigan demostrar, a partir de un número relativamente importante de trabajadores por cuenta ajena, que la retribución media de los trabajadores femeninos es inferior a la de los trabajadores masculinos. Vid., STJUE de 17 de octubre de 1989, asunto 109/88.

sado en datos estadísticos más generales. Estos también sirven para acreditar que «a primera vista [hay] un caso de discriminación» y, por tanto, transferir la carga de la prueba a la parte demandada. Señala el Tribunal de Luxemburgo que, aunque el artículo 19.1 de la Directiva 2006/54/CE no prevé a favor de quien se considera perjudicado un derecho específico de acceso a la información que le permita acreditar hechos en base a los que presumir la existencia de discriminación directa o indirecta, *el hecho de no poder acceder a esa información o a datos estadísticos pertinentes –porque no estén disponibles o son difícilmente accesibles–, en el contexto de la prueba de tales hechos, puede frustrar la consecución del objetivo perseguido por tal Directiva, privando de este modo a dicha disposición de su efecto útil*⁶³⁴. Razón por la cual, teniendo en cuenta la necesidad de garantizar la eficacia del artículo 19.1 de la Directiva 2006/54/CE, esta disposición debe interpretarse en el sentido de que permite que un trabajador que se considere perjudicado por una discriminación indirecta por razón de sexo fundamente una presunción de discriminación basándose en datos estadísticos generales concernientes al mercado de trabajo en el Estado miembro de la UE de que se trate, en el supuesto de que no quepa esperar que el interesado presente datos más concretos relativos al grupo de trabajadores pertinente, por ser difícilmente accesibles o incluso no estar disponibles⁶³⁵.

Pues bien, tanto la doctrina de la falta de transparencia establecida por el TJUE en el caso *Danfoss*, como el enfoque adoptado en relación con la negativa del empresario a proporcionar información a la parte demandante en el caso *Meistery* y la interpretación favorable a aligerar la carga de la prueba en el caso *Schuch-Ghannadan*, podrían ayudar a las víctimas a establecer un caso *prima facie* de discriminación algorítmica. Se trataría, a fin de cuentas, de extrapolar la doctrina ya sentada por el TJUE en dichos casos de discriminación

⁶³⁴ También en el asunto C-104/10 (*Patrick Kelly/National University of Ireland [University College, Dublin]*), el TJUE sostuvo que ni la Directiva relativa a la carga de la prueba en los casos de discriminación por razón de sexo (97/80/CE) ni la Directiva sobre igualdad de trato (76/207/CEE) preveían en general el derecho de los candidatos a una formación profesional a acceder a información sobre la formación de los demás candidatos sobre la base de una sospecha de discriminación, y que toda divulgación debía ajustarse a lo dispuesto en las normas europeas relativas a la confidencialidad de los datos personales. Sin embargo, *correspondía al órgano jurisdiccional nacional dilucidar en cada caso concreto si resultaba necesario divulgar dicha información para cumplir el propósito de la Directiva 97/80/CE* (es decir, para aligerar la carga de la prueba de la discriminación). En dicho caso, el reclamante había solicitado el ingreso en un programa de formación profesional en el University College Dublin (UCD) pero se había rechazado su candidatura y él consideraba estar más cualificado que una candidata a la que se había concedido una plaza. Alegó discriminación por razón de sexo y solicitó que se le exhibieran las demás candidaturas a fin de poder probar los hechos.

⁶³⁵ Como señala CASADESÚS RIPOLL, P., *op.cit.*, p. 139, en estos casos tal presunción no dependerá de las estadísticas ni del grado de probabilidad de que se produzca, sino que cabrá determinar si se ha producido discriminación partiendo de unos hechos determinados, sobre todo cuando el demandado sea poco transparente y se muestre reacio a colaborar.

tradicionales a los casos en que la misma se produce a través de las tecnologías digitales emergentes⁶³⁶.

En tercer lugar, si el proveedor o responsable del despliegue del sistema de IA obstruyera de forma total e injustificada el acceso a la información sobre el mismo, creemos que el juez podría acudir al *principio de disponibilidad y facilidad* probatoria del art. 217.7 LEC, de forma que fuera el demandado el que debiera probar la inexistencia de discriminación si considera que es excesivamente complicado (o incluso imposible) para el demandante conseguir pruebas (o incluso meros indicios) de que el sistema lo ha discriminado⁶³⁷.

Debe tenerse en cuenta que, si bien con carácter general la carga de la prueba supone que demandante y demandado deban probar aquellos hechos en los que basan su acción o defensa respectivamente (*ex* art. 217 apartados 2 y 3⁶³⁸), también se contempla en el apartado 6 del mismo precepto que una disposición legal expresa puede modificar esta regla general, además de que, conforme al último apartado del citado artículo, para su aplicación, el tribu-

⁶³⁶ Concluyen XENIDIS, R./SENDEN, L., «EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination», en Ulf Bernitz et al. (eds.), *General Principles of EU law and the EU Digital Order*, 2020, p. 24, que «se podría argumentar que los requisitos básicos de transparencia, apertura y divulgación deben aplicarse a los modelos algorítmicos para permitir la identificación de prácticas y efectos discriminatorios y permitir a los solicitantes presentar casos *prima facie* de discriminación. Esto es aún más relevante en el contexto de la investigación actual sobre algoritmos auditables, responsabilidad algorítmica e IA explicable, todo ello orientado a permitir una mejor comprensión de cómo funciona la IA, mejorando la interpretabilidad y aumentando la transparencia de la toma de decisiones asistida por IA».

⁶³⁷ Como indica NAVAS NAVARRO, en virtud de dicha regla, correspondería la prueba a aquél que puede acreditar los elementos del caso a un menor coste por hallarse más próximo a las fuentes de prueba o al conocimiento de los hechos y a la disponibilidad de los medios de prueba. En opinión de esta autora, esta regla no se trata de una inversión de la carga de la prueba, sino de lo que se ha denominado como *distribución dinámica de la carga de la prueba*, de suerte que se exige a la parte que dispone de los medios de prueba o que fácilmente puede obtenerlos, que los aporte al proceso para que puedan ser valorados por el juez. En la misma línea, CASADESÚS RIPOLL, P., *op.cit.*, p. 133, afirma que, en base a la jurisprudencia del TS, la regla de la flexibilización de los criterios de atribución de la carga de la prueba no constituye una inversión de dicha carga propiamente dicha. Según la STS 208/2019, de 5 de abril de 2019, el hecho de acudir «al paliativo de la facilidad probatoria, a fin de aliviar al perjudicado de su pesada carga, aminorando el rigor probatorio, no implica una inversión de la carga de la prueba ni una objetivación de la responsabilidad». No obstante, como explica EVANGELIO LLORCA, R., *op.cit.*, p.571, no hay consenso sobre si este principio implica una inversión de la carga de la prueba. *Vid.*, la doctrina citada por la autora defendiendo ambas posturas.

⁶³⁸ «[...] 2. Corresponde al actor y al demandado reconviniendo la carga de probar la certeza de los hechos de los que ordinariamente se desprenda, según las normas jurídicas a ellos aplicables, el efecto jurídico correspondiente a las pretensiones de la demanda y de la reconvencción; 3. Incumbe al demandado y al actor reconvenido la carga de probar los hechos que, conforme a las normas que les sean aplicables, impidan, extingan o enerven la eficacia jurídica de los hechos a que se refiere el apartado anterior».

nal «deberá tener presente la disponibilidad y facilidad probatoria que corresponde a cada una de las partes del litigio». Por tanto, la regla general acerca de la carga de la prueba debe interpretarse teniendo en cuenta los criterios correctores de disponibilidad y facilidad probatoria de cada una de las partes enfrentadas del art. 217.7 LEC, que permiten al juez modificar en el caso concreto las reglas generales sobre su distribución y hacer recaer sobre el demandado la carga de aportar al proceso aquellos elementos que demuestren los hechos que esgrime para su defensa. Así lo ha afirmado el TS en múltiples ocasiones acogiendo la doctrina sentada por el TC: *«Cuando las fuentes de la prueba se encuentran en poder de una de las partes en litigio, la obligación constitucional de colaborar con los tribunales en el curso del proceso (art. 188 de la Constitución) conlleva que dicha parte es quien debe aportar los datos requeridos; que los tribunales no pueden exigir de ninguna de las partes una prueba imposible o diabólica, so pena de causarle indefensión contraria al artículo 24 de la Constitución, por no poder justificar procesalmente sus derechos e intereses pertinentes para su defensa; y que los obstáculos y dificultades puestos por la parte que tiene en su mano acreditar los hechos determinantes del litigio, sin causa que lo justifique, no pueden repercutir en perjuicio de la contraparte, porque a nadie le es lícito beneficiarse de la propia torpeza»*⁶³⁹. La jurisprudencia viene admitiendo la aplicación de este principio de proximidad o facilidad probatoria en supuestos de riesgos extraordinarios, daño desproporcionado o falta de colaboración del causante del daño cuando éste está especialmente obligado a facilitar la explicación del daño por sus circunstancias profesionales o de otra índole⁶⁴⁰.

Además, en relación con el citado principio cuya importancia en el ámbito de la prueba de la discriminación (algorítmica o no) en absoluto es baladí, hay que tener en cuenta los arts. 328 y 329 LEC sobre el deber de exhibición documental entre las partes. Según el primero, «cada parte podrá solicitar de las demás la exhibición de documentos que no se hallen a disposición de ella y que se refieran al objeto del proceso o a la eficacia de los medios de prueba», y en atención al segundo, ante la negativa injustificada a exhibir dichos documentos, «el tribunal, en lugar de lo que en dicho apartado se dispone, podrá formular requerimiento, mediante providencia, para que los documentos cuya exhibición se solicitó sean aportados al proceso, cuando así lo aconsejen las características de dichos documentos, las restantes pruebas aportadas, el contenido de las pretensiones formuladas por la parte solicitante y lo alegado para fundamentarlas».

⁶³⁹ Entre otras, SSTS 684/1999, de 26 de julio; 48/2000, de 31 de enero; 75/2010, de 10 de marzo; 532/2013, de 19 de septiembre.

⁶⁴⁰ STS 141/2021, de 15 de marzo de 2021.

Por último, en virtud de la PDRC, dicha negativa a proporcionar información algorítmica sería un requisito previo para solicitar judicialmente la exhibición de pruebas que obran en poder del demandado o potencial demandado, siempre que dicha solicitud estuviera respaldada por hechos y pruebas suficientes para acreditar la viabilidad de la demanda⁶⁴¹. Dicha medida está recogida en su art. 3 y su objetivo es facilitar la carga de la prueba de la culpa y la causalidad al demandante o potencial demandante⁶⁴².

Esta Propuesta recoge, además, que si el demandado se negara a la solicitud judicial de exhibición de pruebas, se presumiría *iuris tantum* su culpa ex art. 3.5, y presumida la misma (que se equipara con el incumplimiento de los deberes de diligencia a los que le obliga el RIA), se podría presumir posteriormente la relación de causalidad, siempre y cuando pudiese considerarse «razonablemente probable» por el juez, en función de las circunstancias en que se produjo el daño, la existencia de nexo causal entre la culpa y el resultado del sistema⁶⁴³. Ahora bien, no se aplicaría la presunción de causalidad y el demandante debería probarla, cuando el demandado demostrara que el demandante (presunta víctima) pudo acceder razonablemente a pruebas y conocimientos especializados suficientes para demostrar el nexo causal (art.4.4)⁶⁴⁴. Esta excepción a la presunción del nexo causal pretende incentivar a proveedores y responsables del despliegue a facilitar información sobre los sistemas de IA a los perjudica-

⁶⁴¹ Aunque se aclara en el Considerando 17 que «la negativa del proveedor, de la persona sujeta a las obligaciones de un proveedor o del usuario previa a la solicitud de exhibición de pruebas al órgano jurisdiccional no debe dar lugar a la presunción de incumplimiento de las obligaciones de diligencia pertinentes por parte de la persona que deniegue dicha exhibición», según disponen los arts. 3.1 y 3.2, es un requisito previo para que el demandante o potencial demandante pueda solicitar al juez la exhibición de pruebas. Es decir, a tenor de dichos preceptos, para poder hacer uso del derecho a la exhibición de pruebas, el demandante o potencial demandante debe de haber solicitado extrajudicialmente al proveedor, persona sujeta a las obligaciones del proveedor, o al usuario, que le exhiba la información que obra en su poder y éste se haya negado. Además, respecto al demandante, el art. 3.2 dice literalmente que debe haber realizado «todos los intentos proporcionados de obtener del demandado las pruebas pertinentes». No bastará, parece, con solicitar sin éxito las pruebas a las personas mencionadas en el ap.1.

⁶⁴² También se recoge este derecho a la exhibición de pruebas en el art. 9 de la ya aprobada Directiva (UE) 2024/2853 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, sobre responsabilidad por los daños causados por productos defectuosos y por la que se deroga la Directiva 85/374/CEE del Consejo.

⁶⁴³ Será en el Capítulo V, dedicado a las medidas para reparar daños ocasionados por la discriminación algorítmica, donde se analizará con detalle esta Propuesta de Directiva.

⁶⁴⁴ Esto podría ocurrir cuando proveedores o usuarios hubieran colaborado con él extrajudicialmente proporcionando la información solicitada, permitiendo que experimentara con el programa, etc., o cuando el demandante fuera el propio responsable del despliegue del sistema que reclama al proveedor, ya que, en estos casos, debido a las obligaciones de documentación, registro, transparencia y comunicación de información que el proveedor debe cumplir respecto del responsable, éste contaría ya con la información del sistema.

dos, así como a cumplir las medidas del RIA dirigidas a garantizar un alto nivel de transparencia y el cumplimiento de requisitos de documentación y registro.

Por lo que se refiere a los casos de discriminación algorítmica, se podría plantear la duda de si, para aligerar al demandante la carga de probar la discriminación, tendrían que aplicarse las presunciones refutables de los arts. 3 y 4 de la Propuesta de Directiva citada, o si podría aplicarse la inversión de la carga de la prueba recogida tanto en las Directivas de igualdad como en las normas nacionales de transposición⁶⁴⁵. Al respecto considera NAVAS NAVARRO, cuya opinión compartimos, que la respuesta podría estar en el art. 1.4 de la Propuesta⁶⁴⁶. Esto es, si el enfoque de armonización mínima que adopta la Propuesta de Directiva permite que los demandantes puedan invocar normas del Derecho nacional más favorables que las previstas en ella, no cabe duda de que en materia de no discriminación la norma más beneficiosa vendría a ser la que instaura la inversión de la carga de la prueba (art. 30 Ley 15/2022 integral de igualdad en relación con el art. 217.5 LEC, entre otros), norma que además es compatible con el Derecho de la Unión, tal como exige el precepto mentado.

3.1.4. Justificación de la discriminación algorítmica

Una vez acreditada la existencia de indicios razonables de discriminación y trasladada la carga de la prueba al demandado, habría que preguntarse qué debería probar éste para no responder.

Según la doctrina constitucional, «cuando el recurrente alega una discriminación prohibida por el art. 14 CE [...], aportando indicios racionales de discriminación, corresponde a la empleadora la obligación de rebatirlos *justificando que su actuación fue absolutamente ajena a todo propósito atentatorio de derechos fundamentales* (SSTC 17/2007, de 12 de febrero, FJ 4, y 173/2013, de 10 de octubre, FJ 6, entre otras). Pero, incluso si dicha intencionalidad discriminatoria no existe, corresponde también al empleador *probar que la vulneración que se le atribuye no represente objetivamente actos contrarios a la prohibición de discriminación* (en este sentido, STC 233/2007, de 5 de noviembre, FJ 4)»⁶⁴⁷. Por tanto, «una vez que quien invoca la aplicación de la regla de la prueba indiciaria desarrolla una actividad alegatoria suficientemente precisa y concreta en torno a los indicios de la existencia de discriminación, recaerá sobre la parte

⁶⁴⁵ Aunque suele hablarse de *inversión de la carga de la prueba*, en puridad se trata más bien de un *reparto de la carga de la prueba* como explicamos *supra*.

⁶⁴⁶ *Vid.*, «Régimen europeo en ciernes en materia de responsabilidad derivada de los Sistemas de Inteligencia Artificial», *Revista CESCO de Derecho de Consumo*, n. 44, 2022, p. 35.

⁶⁴⁷ STC 51/2021, de 15 de marzo.

demandada la carga de probar la existencia de causas suficientes, reales y serias, para calificar de razonable y ajena a todo propósito lesivo del derecho fundamental la decisión o práctica empresarial cuestionada, único medio de destruir la apariencia lesiva creada por tales indicios (STC 2/2009, de 12 de enero, FJ 3)” (STC 31/2014, FJ 3)»⁶⁴⁸.

La Agencia de los Derechos Fundamentales de la Unión Europea afirma que en los casos de discriminación *directa* el presunto autor de la discriminación podría destruir esa presunción de dos formas: o bien probando que el reclamante no se encuentra en situación similar a su «referencia comparativa», o que la diferencia de trato no se basa en el motivo protegido, sino en otras diferencias objetivas. Si el presunto autor no puede destruir la presunción, tendrá que aportar una causa justificativa de la diferencia de trato y demostrar que se trata de una medida objetivamente justificada y proporcionada⁶⁴⁹. Y en los casos de discriminación *indirecta*, deberá probar que la medida, práctica o conducta aparentemente neutra –pero que tenía efectos perjudiciales para el grupo protegido– perseguía una finalidad legítima, que los medios para conseguirla eran adecuados y necesarios, y que las desventajas causadas para lograr los objetivos perseguidos no eran desproporcionadas, cuestiones que serán sometidas a consideración o valoración del tribunal⁶⁵⁰.

El triple juicio de proporcionalidad de la medida discriminatoria que debe realizar el tribunal implica la valoración de tres extremos: en primer lugar, si la medida era adecuada a los fines perseguidos; en segundo lugar, si era necesaria, en el sentido de que no existiera ninguna otra medida menos gravosa para conseguir dichos fines; y en tercer lugar, si era proporcionada en sentido estricto, esto es, si finalidad perseguida era suficientemente importante como para justificar ese grado de injerencia.

Respecto a los casos de discriminación algorítmica se ha mantenido que para que el sistema de IA pudiera superar ese triple juicio de proporcionalidad que debe realizar el juez cuando el demandado trata de demostrar que la

⁶⁴⁸ STC 31/2014, de 24 de febrero. Se añade, además, que la carga de demostrar el carácter justificado de la diferenciación «se torna aún más rigurosa que en aquellos casos que quedan genéricamente dentro de la cláusula general de igualdad del art. 14 CE, al venir dado el factor diferencial por uno de los típicos que el art. 14 CE concreta para vetar que en ellos pueda basarse la diferenciación, como ocurre con el sexo, la raza, la religión, el nacimiento y las opiniones» (STC 81/1982, de 21 de diciembre).

⁶⁴⁹ Con arreglo al CEDH, resulta posible recurrir al criterio de justificación por motivos objetivos, mientras que, en el Derecho de la Unión, la diferencia de trato únicamente puede justificarse en determinados casos. Así se explica en la p. 269 del *Manual de legislación europea contra la discriminación*, ya citado.

⁶⁵⁰ *Manual de legislación europea contra la discriminación*, cit., p. 270. Según el TJUE, debe entenderse por «finalidad legítima» cualquier finalidad lícita que no exprese en sí misma un ánimo de discriminar o no trate de ocultarlo de forma evidente.

medida, decisión o práctica adoptada es adecuada, objetiva y razonable, habría que probar que el sistema era suficientemente preciso para predecir el fenómeno para el que se creó, que su diseño era suficientemente transparente para permitir su vigilancia y control a lo largo de su ciclo de vida y que los datos empleados eran de calidad⁶⁵¹.

También se ha dicho que al responsable o al proveedor del sistema le correspondería probar que no contiene errores, que no comporta sesgos, que la diferencia de trato se justifica en un elemento objetivo y no en un motivo prohibido, que el tratamiento diferente que genera el sistema tiene una justificación objetiva y razonable, suficientemente probada, de las medidas adoptadas y de su proporcionalidad, etc. El proveedor, y en su caso el responsable del despliegue, deberían indicar la finalidad pretendida y cómo se ha configurado el sistema, la selección de datos de entrada y entrenamiento y otros elementos, y que estas decisiones técnicas y organizativas arrojan resultados proporcionales respecto de tales finalidades y con relación a las diferencias que generan⁶⁵².

A tenor de lo establecido en la PDRC, parece que cuando se utilicen sistemas de IA para tomar decisiones, los proveedores y/o usuarios del sistema (responsables del despliegue en los términos definitivos del RIA) que presuntamente arroja resultados discriminatorios, deberán acreditar el cumplimiento de las obligaciones contempladas en los arts. 4.2 y 4.3 respectivamente, para exonerarse de responsabilidad.

Así, si el demandado es el *proveedor de un sistema de alto riesgo*, debería probar que el sistema IA: a) Se ha desarrollado a partir de conjuntos de datos de entrenamiento, validación y prueba que cumplan los criterios de calidad del art. 10 Ley de IA; b) Ha sido diseñado y desarrollado de modo que cumpla

⁶⁵¹ Respecto a la prueba de que no había otro sistema de IA menos discriminatorio para conseguir los fines previstos, mantienen GERARDS, J./XENIDIS, R., *op.cit.*, p. 73, que será un desafío para los jueces valorar si una empresa de *software* responsable del resultado discriminatorio de un algoritmo podría haber optado por una solución menos perjudicial a la luz de las explicaciones altamente técnicas que probablemente presente en relación con los costes o sacrificios que debe hacer. Sobre este asunto considera también SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación...», *cit.*, p. 13/31, que «con el objetivo de no imponer costes inasumibles a algunos operadores tecnológicos, deberá exigírseles el uso de sistemas no discriminatorios dentro del criterio de “las mejores técnicas disponibles”, esto es, extrapolando dicha figura aplicada en el ámbito del derecho medioambiental a los riesgos generados por las nuevas tecnologías de procesamiento de datos. Es decir, deberá tenerse en cuenta tanto el estado de la técnica como la capacidad económica de la parte demandada».

⁶⁵² VELASCO RICO, C.L. «Marco regulatorio de los sistemas algorítmicos y de IA. El papel de la Administración», ponencia en el XVIII Congreso de la Asociación Española de Profesores de Derecho Administrativo: El derecho administrativo en la era de la inteligencia artificial, Vigo 25 a 27 enero de 2024, p. 22.

con los requisitos de transparencia del art. 13 Ley de IA; c) Ha sido diseñado y desarrollado de modo que permita su vigilancia efectiva por personas físicas en el periodo en que se esté utilizando (art. 14 Ley de IA); d) Ha sido diseñado y desarrollado de modo que a la luz de su finalidad prevista, alcance un nivel adecuado de precisión, solidez y ciberseguridad (art. 15 y art. 16 letra a Ley de IA); e) Se han adoptado de forma inmediata las medidas correctoras necesarias para poner el sistema de IA en conformidad con los deberes de cuidado establecidos en el [título III, capítulo 2, de la Ley de IA] o para retirar el sistema del mercado o recuperar el sistema según proceda, de conformidad con la Ley de IA (art. 16 letra g y art. 21 Ley de IA).

En caso de que el demandado sea el usuario del sistema de alto riesgo, debería probar *ex art. 4.3*: a) Que cumplió con sus obligaciones de utilizar o supervisar el sistema de IA de conformidad con las instrucciones de uso adjuntas o, en su caso, de suspender o interrumpir su uso (art. 29 de la Ley de IA); y b) Que expuso al sistema de IA a datos de entrada bajo su control que eran pertinentes habida cuenta de la finalidad prevista del sistema (art. 29.3 de la Ley IA). No obstante, el Considerando 26 de la Propuesta señala que la culpa del usuario de sistemas de IA de alto riesgo podría determinarse también acreditando el incumplimiento de otras obligaciones de diligencia establecidas en el Derecho de la Unión o nacional según establece el art. 29.2 Ley IA⁶⁵³.

Ahora bien, la prueba del cumplimiento de todos estos parámetros de seguridad, ¿implicaría necesariamente descartar que el sistema discrimina? Es decir, ¿podría ocurrir que aun habiendo cumplido todos los requisitos y obligaciones que incumben a los operadores mencionados, el sistema de IA arrojará resultados discriminatorios?

A nuestro modo de ver, si bien esta posibilidad será difícil teniendo en cuenta las numerosas exigencias impuestas por el RIA para los sistemas de alto riesgo, no es descartable. Es decir, podríamos encontrarnos con sistemas altamente precisos, sólidos, seguros y transparentes (o que cumplieran ese «nivel adecuado de precisión, solidez y ciberseguridad» al que hacen referencia las propuestas europeas, así como el resto de exigencias mencionadas), pero cuyos resultados –fruto del aprendizaje autónomo del sistema– tendie-

⁶⁵³ Para MARTÍ GRAU, R., «Reflexiones acerca de la Propuesta de Directiva sobre responsabilidad por daños derivados de la inteligencia artificial y su impacto en el Derecho español de daños», *Revista Aranzadi Doctrinal*, num.4/2023, BIB 2023\705, p. 4/14, los casos de culpa del demandado previstos en los apartados segundo y tercero del art. 4, no excluyen la apreciación de la infracción de normas de cuidado de acuerdo con el Derecho nacional, incluyendo las normativas sectoriales que en su caso resulten de aplicación. Y para EVANGELIO LLORCA, R., *op.cit.*, p. 605, teniendo en cuenta que la proyectada Directiva es de mínimos, nada impediría que los legisladores nacionales ampliaran el catálogo de faltas de diligencia previsto en aquella.

ran a perjudicar a personas o grupos protegidos⁶⁵⁴. En tales casos, al tratarse de una responsabilidad subjetiva o por culpa, si proveedores y/o responsables del despliegue probaran el cumplimiento de las obligaciones anteriores quedarían exonerados de responsabilidad. Y ello aun cuando el sistema arrojará resultados discriminatorios. Por tanto, conforme a la regulación propuesta en esta Directiva, parece que prevalecerían los intereses empresariales y el uso de la tecnología sobre los derechos de igualdad y no discriminación. Es decir, la prueba de la diligencia debida justificaría ese grado de inferencia en dichos derechos.

No obstante, aunque no quedaran obligados a responder por el daño en base a dicha normativa, si el sistema de IA discriminara podría exigirse responsabilidad en base a la normativa antidiscriminatoria y/o a la de protección de datos personales cuando se probase su incumplimiento, ya que, como se afirma en la Exposición de Motivos de la Directiva proyectada, la misma es complementaria de otras políticas de IA de la Comisión como el RGPD y la legislación de la UE sobre no discriminación e igualdad de trato, entre otras⁶⁵⁵. Al complementar estas otras vertientes, la propuesta protege el derecho de la víctima a una indemnización en virtud del Derecho privado, incluidas las indemnizaciones por violaciones de los derechos fundamentales.

En el caso de que los proveedores y/o usuarios no consiguieran demostrar que cumplieron las obligaciones anteriormente mencionadas, se entenderá probada su culpa, y según el art.4 PDRC, la prueba de la culpa dará lugar a la presunción de la relación de causalidad entre ésta y el resultado dañoso del sistema, siempre y cuando pudiera considerarse «razonablemente probable» por el juez, en función de las circunstancias en que se produjo el daño, la existencia de dicho nexo causal. Sobre esta cuestión volveremos en el capítulo V.

3.2. BREVE EXPLICACIÓN DE LA PROBLEMÁTICA QUE PRESENTA LA FIJACIÓN DEL NEXO CAUSAL EN EL CASO DE LOS DAÑOS OCASIONADOS POR SISTEMAS DE IA: LA DIFICULTAD DE SU PRUEBA Y LOS MECANISMOS PARA MITIGARLA

⁶⁵⁴ SORIANO ARNANZ, A., «La aplicación del marco jurídico europeo ...», cit., p. 78, pone como ejemplo el caso de un sistema de selección laboral que elige a las mejores candidatas con una precisión del 95% pero del que se demuestra que sus resultados suelen perjudicar en mayor medida a las mujeres que a los hombres.

⁶⁵⁵ Será en el Capítulo V cuando se aborden las medidas previstas en dichos cuerpos normativos tendentes a conseguir la reparación de los daños derivados de la discriminación.

Como explicamos en su momento, la discriminación algorítmica puede provenir de diversas causas o factores. Puede tener su raíz en el propio diseño del sistema, en los datos utilizados, en las inferencias a las que llega en base a su aprendizaje autónomo o en otras causas como la interpretación humana de sus resultados. Además, el origen de la discriminación puede ubicarse en las diferentes etapas del proceso de toma de decisión.

Si bien la determinación del nexo de causalidad entre la conducta de un sujeto y el daño ha sido un problema clásico en el ámbito de la responsabilidad civil, la cuestión se complica considerablemente cuando entran en juego los sistemas de IA. En el ámbito algorítmico no va a ser tarea fácil saber hasta qué punto un daño es resultado del propio aprendizaje autónomo del sistema de IA, o si la causa del daño se sitúa en un momento anterior y proviene de la acción u omisión de las personas que intervienen en su creación, desarrollo o implementación, que han errado, por ejemplo, en la elección del tipo de algoritmo para alcanzar el objetivo pretendido, en las variables introducidas o en la selección de los datos de entrenamiento⁶⁵⁶. Esto dificulta a su vez la averiguación de los sujetos responsables de los daños, problema que abordaremos después.

Estas pesquisas se complican cuando se emplean algoritmos de aprendizaje profundo basados en redes neuronales que funcionan a modo de caja negra, impidiendo llegar a saber por qué han decidido como lo han hecho⁶⁵⁷. Además, incluso si la fuente del problema se pudiera llegar a identificar, estos algoritmos están en constante evolución y, en consecuencia, la causa u origen

⁶⁵⁶ Entre las causas que pueden provocar los daños ocasionados por los sistemas de IA, la doctrina menciona los siguientes: datos sesgados o insuficientes, defectos de conectividad como fallos de los dispositivos interconectados o de la conexión a la red u otras infraestructuras o instalaciones técnicas que forman parte del funcionamiento del sistema de IA, así como fallos de los sensores internos o de los algoritmos incorporados diseñados para analizar, verificar y procesar dichos datos. Por todos, EVANGELIO LLORCA, R., *op.cit.*, p. 562 y doctrina citada.

⁶⁵⁷ Como explica EBERS, M., «La utilización de agentes electrónicos inteligentes en el tráfico jurídico: ¿Necesitamos reglas especiales en el Derecho de la responsabilidad civil?», *InDret*, n. 3, 2016, p. 12, que el error del sistema no pueda ser verificado *ex post* depende justamente del funcionamiento de las redes artificiales neuronales: «Mientras que en la IA simbólica, el programa puede examinarse línea por línea, instrucción por instrucción, en las redes artificiales neuronales desaparece la representación simbólica del saber y el control del proceso. En lugar de símbolos claros, comprensibles, contamos con una matriz de pesos sinápticos que se sustraen de la interpretación directa. El saber y el comportamiento almacenados en la red neuronal tan solo se revelan –como en los organismos vivos– a través de averiguaciones indirectas». Precisamente por esto, añade ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...cit.*, pp. 173-174, que de poco o nada sirve en estos últimos sistemas investigar las cajas negras que en cambio sí que podrían aportar información relevante cuando se trata de algoritmos deterministas.

de la discriminación podría cambiar con el tiempo o desaparecer por completo, no habiendo entonces ninguna evidencia para demostrarla⁶⁵⁸.

Junto a la complejidad y opacidad de los sistemas, su interconectividad y dependencia de datos también plantean problemas en la lógica tradicional de la responsabilidad civil, puesto que conducen a una sobredeterminación de causas que complica la identificación de las personas que pueden haber intervenido en la producción del daño⁶⁵⁹. Además, la cada vez mayor autonomía de los sistemas complica que la responsabilidad pueda imputarse a una conducta o actividad *humana*, tal como exige el Derecho de daños⁶⁶⁰.

Las dificultades encontradas para probar tanto el concreto defecto de los sistemas de IA, como la relación causal entre éste y el daño ocasionado «más allá de toda duda razonable»⁶⁶¹, llevaron tanto a las instituciones europeas como a la doctrina a proponer mecanismos para mitigarla. Entre ellos, por ejemplo, aplicar la «sospecha de defecto» que puede relacionarse con la «causalidad probabilística» que se aplica cuando no hay certeza acerca del nexo causal⁶⁶², exonerar a la víctima de probar la parte concreta del producto que resulta defectuosa bastando con acreditar que en su conjunto tiene un defecto, obligar a los desarrolladores del *software* a registrar información suficiente

⁶⁵⁸ GERARDS, J./XENIDIS, R., *op.cit.*, pp. 73 y ss.

⁶⁵⁹ EVANGELIO LLORCA, R., *op.cit.*, p. 562 y NAVAS NAVARRO, S., «Sistemas expertos ...», *cit.*, p. 4/16.

⁶⁶⁰ Como advierte EVANGELIO LLORCA, R., *op.cit.*, p. 566, en muchos casos los daños causados por estos sistemas solo pueden atribuirse de forma muy remota e indirecta a un comportamiento humano, precisamente porque son capaces de tomar decisiones concretas conforme a parámetros generales que interpretan de forma autónoma.

⁶⁶¹ Este estándar de prueba es el que se exige en la mayoría de ordenamientos de Derecho continental y en ocasiones se expresa como una «probabilidad rayana en la certeza», lo que significa que el juez debe estar convencido más allá de toda duda razonable. Explica EVANGELIO LLORCA, R., *op.cit.*, pp. 575, que si hablamos del nexo causal, el TS considera que sería preciso una prueba terminante acerca del vínculo entre la acción u omisión del agente y el daño para declarar la responsabilidad, no siendo suficientes las simples conjeturas, o la existencia de datos fácticos que, por una mera coincidencia, induzcan a pensar en una posible interrelación de esos acontecimientos. No obstante, sigue diciendo esta autora que «existe amplio consenso en que no es razonable mantener un estándar de prueba tan alto en determinados ámbitos en los que al perjudicado le puede resultar particularmente costoso averiguar la causa del daño, por diversas razones, entre las cuales se suelen mencionar las siguientes: sectores en los que el desarrollo científico y tecnológico es particularmente intenso, ignorándose en muchos casos cómo se desenvuelven con exactitud los procesos causales en las aplicaciones científicas y tecnológicas; difícil acceso a los medios, materiales o documentos probatorios; etc».

⁶⁶² La causalidad probabilística es un importante expediente para rebajar el estándar de prueba exigible al que han recurrido los tribunales españoles en ámbitos donde es muy complicado probar los presupuestos de la responsabilidad civil, como en materia de daños causados por productos defectuosos. En virtud del mismo, «se da por probada la causa cuando es *probable* que haya sido la que ha irrogado el daño, sustituyéndose la prueba de la causalidad física por la de la *probabilidad* de que se haya causado el daño». Así, EVANGELIO LLORCA, R., *op.cit.*, p. 576.

sobre su funcionamiento y presumir la causalidad en caso de no hacerlo, o invertir la carga de la prueba⁶⁶³.

En su *Informe sobre la aplicación de la Directiva en materia de responsabilidad por daños causados por productos defectuosos*, la Comisión Europea ya reconocía que con carácter general para cualquier tipo de producto (y no solo en los casos de daños causados por la IA), lo más complicado para los perjudicados era probar la relación de causalidad⁶⁶⁴. No obstante, el Grupo de Expertos en responsabilidad y nuevas tecnologías también reconocía en su *Informe de 21 noviembre 2019 sobre responsabilidad de la IA y otras tecnologías emergentes*, que cuanto más compleja fuera la interacción de los factores que contribuyen al daño (como ocurre en los sistemas inteligentes en los que interaccionan múltiples

⁶⁶³ Todos estos mecanismos son explicados por ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...*cit., pp. 175-181. Para esta autora, la medida que implica exonerar a la víctima de tener que probar la parte concreta del producto que resulta defectuosa, siendo suficiente con acreditar que en su conjunto tiene un defecto, es especialmente interesante en los sistemas inteligentes, puesto que la interacción de múltiples elementos físicos y digitales dificulta dicha prueba. Con ella se evitarían los problemas para determinar el sujeto responsable, ya que todos aquellos que hubieran aportado elementos físicos o digitales a dicho producto deberían responder solidariamente en cuanto fabricantes del mismo. Vid., p. 177. Esta solución ya se proponía en el Informe del Grupo de Expertos en responsabilidad y nuevas tecnologías, de 21 noviembre 2019, sobre responsabilidad de la IA y otras tecnologías emergentes y se recogió en la respuesta del ELI a la consulta pública sobre la adaptación de las reglas sobre responsabilidad civil a la era digital y la IA lanzada por la Comisión. Nótese también que, con carácter general en el ámbito de la responsabilidad civil por productos defectuosos, se ha venido manteniendo por doctrina y jurisprudencia que el hecho de que perjudicado tuviera que probar el defecto del producto no significaba que tuviera que identificar la causa o el origen del mismo. En relación con el art. 139 TRLGDCU (o su precedente, el art. 5 Ley 22/1994), según el cual «[e]l perjudicado que pretenda obtener la reparación de los daños causados tendrá que probar el defecto, el daño y la relación de causalidad entre ambos», es doctrina reiterada del TS español la de que incumbe al actor demostrar que el producto es defectuoso, pero no el concreto defecto. Vid. STS 332/2008, de 30 de abril, ES:TS:2008:1545.

⁶⁶⁴ Vid., *Informe de la Comisión al Parlamento Europeo, al Consejo y al Comité Económico y Social Europeo sobre la aplicación de la Directiva del Consejo relativa a la aproximación de las disposiciones legales, reglamentarias y administrativas de los Estados miembros en materia de responsabilidad por los daños causados por productos defectuosos* (85/374/CEE). COM (2018) 246 final, p. 5. También manifiestan GOMEZ LIGUERRE, C./GARCÍA-MICO, T., «Liability for Artificial Intelligence and other emerging technologies», *Indret*, n.1, 2020, pp. 502-503, que en los pleitos actuales en materia de responsabilidad por productos, el 53% de los casos que se desestiman lo son por falta de prueba suficiente del defecto en el producto o del nexo causal con el daño. Y NAVAS NAVARRO, S., «Responsabilidad civil del fabricante y tecnología inteligente. Una mirada al futuro», *Diario La Ley*, N° 35, Sección Ciberderecho, 27 de Diciembre de 2019, LA LEY 15349/2019, p. 8/16, relata que ya en el año 1999, con ocasión del Informe presentado por la Comisión sobre la responsabilidad civil por productos defectuosos, se pusieron de manifiesto los problemas relacionados con la prueba del defecto y nexo causal y se plantearon posibles alternativas a la legislación.

elementos físicos y digitales), más difícil sería para la víctima tener éxito en el establecimiento de la causalidad sin aliviar su carga de la prueba⁶⁶⁵.

Para paliar este problema, y sobre la base de que la complejidad de esta tecnología incrementaba la asimetría informativa entre el responsable y su víctima hasta el punto de convertir la prueba del vínculo causal en imposible o excesivamente costosa, se proponían en dicho Informe una serie de medidas para facilitar la carga probatoria del reclamante. Se recogían una serie de supuestos que, valorados en su conjunto, podían justificar que el legislador europeo o los nacionales incluyeran una regla general de inversión de la carga de la prueba de la causalidad. A saber: a) la probabilidad de que la tecnología hubiera contribuido a ocasionar el daño; b) la probabilidad de que el daño se hubiera ocasionado por la intervención de la tecnología o por otra causa que se encontrara dentro de la misma esfera de control; c) el riesgo de que exista un defecto conocido en la tecnología, aunque su impacto en el nexo causal (como causa real) no fuera evidente por sí mismo; d) el grado de trazabilidad *ex post* y de inteligibilidad de los procesos gobernados por IA que pudieran haber contribuido a causar el daño (asimetría informativa); e) el grado de acceso posterior y comprensión de los datos recogidos y generados por la tecnología; y f) el tipo y nivel de daño potencial y del realmente ocasionado.

También se proponía en dicho Informe la posibilidad de presumir la causalidad (además de la culpa y la propia existencia del defecto) cuando se incumplieran normas de seguridad y ciberseguridad cuya observancia podría haber evitado el daño. Como señala ATIENZA NAVARRO, esta circunstancia era especialmente importante, ya que de aplicarse dicha presunción sería suficiente con que la víctima demostrase que el sistema no cumplía con el estándar de seguridad exigible o esperable para que se pudiese presumir la relación de causalidad⁶⁶⁶. Se trataría de una presunción *iuris tantum* en tanto en cuanto los fabricantes u operadores podrían liberarse de responsabilidad probando que aquellas normas se cumplieron, es decir, que se adoptaron todas las medidas de seguridad exigidas.

⁶⁶⁵ Vid., p. 20. En este informe se parte de que «Cuando una tecnología concreta aumente las dificultades para probar la existencia de un elemento de responsabilidad más allá de lo que pueda esperarse razonablemente, las víctimas deben tener derecho a que se facilite la prueba». En la Exposición de Motivos de la PDPD de 2022, se vuelve a poner de manifiesto que esa prueba puede resultar muy complicada en el caso de daños causados por sistemas de IA. En este sentido, el 77% de los encuestados en la encuesta pública llevada a cabo por la CE con vistas a la elaboración de la propuesta de una nueva Directiva en la materia consideró que los productos técnicamente complejos crean dificultades con respecto a la carga de la prueba de la persona perjudicada; siendo el porcentaje considerablemente más elevado entre las organizaciones de consumidores, las ONG y los ciudadanos (95%) que entre las organizaciones empresariales e industriales (38%).

⁶⁶⁶ ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...*cit., p. 180.

Además, para aliviar la carga de la prueba del defecto y la causalidad, el Informe también sugería incorporar la obligación del fabricante o productor de monitorear o hacer un seguimiento de los sistemas de IA una vez puestos en circulación. Se proponía que los productores tuvieran la obligación de dotar a la tecnología de medios para registrar información sobre su funcionamiento (lo que se conoce como registro por diseño o «*logging by design*») que permitiera recabar pruebas de posibles fallos o defectos del sistema. Además, se consideraba necesario que dicha información se pusiera a disposición de los perjudicados para ayudar a esclarecer la responsabilidad, evitando cualquier encriptación⁶⁶⁷. En caso de no facilitarse, o de ausencia de información registrada que impidiera identificar a las personas que manipularon el sistema, se presumiría *iuris tantum* la causalidad, invirtiéndose la carga de la prueba⁶⁶⁸.

A pesar de todas las recomendaciones incluidas en dicho Informe, la posterior Propuesta de Reglamento que acompañaba a la *Resolución del Parlamento de 20 de octubre de 2020, sobre un régimen de responsabilidad civil en materia de inteligencia artificial*, no recogió ninguna presunción o regla de inversión de la carga de la prueba del nexo causal o defecto del sistema de IA. Sólo establecía la inversión de la carga de la prueba de la *culpa* cuando los daños hubiesen sido causados por sistemas de IA que no fueran de alto riesgo, para los que se proponía un régimen de responsabilidad subjetiva.

Por el contrario, las últimas propuestas de Directiva relacionadas con la responsabilidad civil y la IA, de 28 de septiembre de 2022, fueron precisamente en la línea sentada en aquel Informe. Esto es, el establecimiento de presunciones *iuris tantum* tanto de la culpa y la causalidad, como del propio defecto del sistema, cuando se cumplan ciertos requisitos. Tal como explica la propia Comisión, se opta por las presunciones *iuris tantum* por ser la opción menos gravosa o herramienta menos intervencionista para los Estados que, aun así,

⁶⁶⁷ Ahora bien, estamos de acuerdo con ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...*cit., p. 178 y NAVAS NAVARRO, S., «Sistemas expertos basados en inteligencia artificial y responsabilidad civil...», cit., en que dicha solución no será la panacea en tanto que exige elevados conocimientos técnicos para interpretar la información proporcionada, de los que, como regla general, carecerá el ciudadano medio y que obligarán a contar con peritos especializados que encarecerán estas reclamaciones llegando a afectar, de facto, al derecho fundamental a la tutela judicial efectiva. Por ello, la primera autora citada califica esta solución como un mero «parche» que no soluciona realmente el problema.

⁶⁶⁸ Señala VÁZQUEZ DE CASTRO, E., «Aproximación a la responsabilidad derivada de los riesgos de la inteligencia artificial en Europa», en *Dimensiones éticas y jurídicas de la inteligencia artificial en el marco del Estado de Derecho* J.I., Solar Cayón (ed. lit.), 2020, p. 244, que la inexistencia de estos registros o de monitorización del sistema también daría derecho de regreso al empresario que hubiera tenido que responder (como responsable solidario) por haber explotado económicamente el producto y no se hubiera podido librar de responsabilidad por aplicársele la presunción o al carecer de medios de prueba para destruir dicha presunción por no existir registro.

consigue incentivar el cumplimiento de las obligaciones de diligencia en vigor establecidas a escala nacional o de la Unión. Se descarta el establecimiento de una inversión de la carga de la prueba para evitar exponer a proveedores, operadores y usuarios de sistemas de IA a mayores riesgos de responsabilidad civil, ya que esto podría obstaculizar la innovación y reducir la adopción de productos y servicios basados en la IA.

Por lo que se refiere a las obligaciones de monitorización y registro de información sobre el funcionamiento de los sistemas recomendadas en el Informe del Grupo de Expertos, se han recogido en el RIA para los sistemas de alto riesgo. Con ellas se pretende favorecer un nivel de trazabilidad y transparencia adecuado a su finalidad durante todo su ciclo de vida. Así lo dispone el art. 12 donde se establece que los sistemas de IA de alto riesgo *permitirán técnicamente* el registro automático de eventos («archivos de registro») a lo largo de todo su ciclo de vida. Para garantizar un nivel de trazabilidad del funcionamiento del sistema de IA de alto riesgo que resulte adecuado para la finalidad prevista del sistema, las capacidades de registro permitirán que se registren eventos pertinentes para: a) la detección de situaciones que puedan dar lugar a que el sistema de IA de alto riesgo presente un riesgo en el sentido del artículo 79, apartado 1, o a una modificación sustancial; b) la facilitación de la vigilancia poscomercialización a que se refiere el artículo 72⁶⁶⁹; y c) la vigilancia del funcionamiento de los sistemas de IA de alto riesgo a que se refiere el artículo 26, apartado 6. En el caso de los sistemas de IA de alto riesgo que figuren en el punto 1, letra a), del anexo III, las capacidades de registro incluirán, como mínimo: a) un registro del período de cada uso del sistema (la fecha y la hora de inicio y la fecha y la hora de finalización de cada uso); b) la base de datos de referencia con la que el sistema ha cotejado los datos de entrada; c) los datos de entrada con los que la búsqueda ha arrojado una correspondencia; d) la identificación de las personas físicas implicadas en la verificación de los resultados que se mencionan en el artículo 14, apartado 5».

La importancia de tales obligaciones no es baladí teniendo en cuenta que el comportamiento futuro de un algoritmo puede ser difícil de prever al depender en gran medida de los datos con los que interactúa, el ecosistema de retroalimentación que generan esos datos y factores cambiantes relacionados con el contexto socioeconómico o medioambiental. Por ello, es recomendable tanto guardar cada versión del programa utilizado en los procesos de toma de decisiones, como monitorizar en el tiempo el comportamiento del sistema respecto a factores cuya variación pueda afectar a su funcionamiento en el futuro. Estos cambios pueden relacionarse con aquellas variables cuya repre-

⁶⁶⁹ El RIA ha terminado por establecer estas obligaciones de vigilancia poscomercialización en el capítulo IX, sección 1, art. 72.

sentación en las bases de datos de entrenamiento del sistema es demasiado pequeña o demasiado grande, sin que esto se corresponda con la realidad social o sabiendo que la misma puede cambiar, o también pueden afectar a la variable objetivo del modelo alterando su eficiencia. Por ejemplo, si en una localidad donde se está implementando un algoritmo de evaluación de riesgos se produce un notable incremento de población femenina y el algoritmo no es capaz de captar esta transformación social, podría infravalorar el riesgo de las mujeres y limitar los recursos a ellas asignados⁶⁷⁰.

4. TERCER OBSTÁCULO: DETERMINAR LOS SUJETOS RESPONSABLES

4.1. PROBLEMAS ENCONTRADOS PARA LA ATRIBUCIÓN DE RESPONSABILIDAD

Como es lógico, la dificultad para determinar la causa o causas de los daños ocasionados por sistemas de IA repercute inexorablemente en la determinación de los posibles sujetos responsables. Piénsese que, en el complejo entorno de la IA, el *big data* y el *IoT*, la toma de decisiones puede implicar situaciones en las que el resultado de un algoritmo, que a su vez se basa en los datos generados por un dispositivo conectado –por ejemplo, un asistente de voz o un robot de cocina–, se utilice como entrada para otro algoritmo. Esta situación genera múltiples riesgos y problemas, ya que, si uno de los sistemas interconectados falla por un defecto técnico o una mala interpretación de los datos, puede conllevar que otro sistema también falle, dando lugar a un proceso de errores en cascada⁶⁷¹.

Adicionalmente, la multiplicidad de personas físicas o jurídicas que intervienen en el desarrollo de dichos sistemas inteligentes (diseñadores, programadores, entrenadores del algoritmo, terceros que proporcionan datos, analistas de datos, vendedores del *software*, importadores, responsables de su despliegue, encargados de suministrar actualizaciones, etc.) hace que la *atri-*

⁶⁷⁰ ETICAS CONSULTING, *Guía de auditoría algorítmica*, pp. 71-72.

⁶⁷¹ GERARDS, J./XENIDIS, R., *op.cit.*, p. 73. En sentido similar, MARTÍN-CASALS, M., «Las propuestas de la Unión Europea para regular la responsabilidad civil por los daños causados por sistemas de inteligencia artificial», *InDret*, n.3, 2023, p. 67, afirma que «la interconectividad plantea nuevos retos de seguridad y protección frente a daños puesto que será más difícil identificar quiénes son los actores que han intervenido en la producción del daño o si la causa del daño ha sido el suministro por un tercero de datos sesgados, un fallo de alguno de los dispositivos conectados, un fallo de la conexión a la red o de otras infraestructuras o instalaciones técnicas que forman parte del funcionamiento del sistema de IA».

bución y distribución de la responsabilidad sea todavía más complicada. Por ello, aun cuando la discriminación hubiera quedado probada, es un verdadero desafío identificar tanto su causa como los posibles agentes responsables entre la pluralidad de operadores de la cadena de suministro.

Esta problemática ha sido puesta de manifiesto por las instituciones europeas en muchos de sus documentos relacionados con la IA. Entre los más recientes, en la Propuesta de Directiva sobre responsabilidad civil extracontractual de la Comisión Europea⁶⁷². También en la Orientación General del Consejo UE sobre la Propuesta de RIA, se consideró necesario introducir algunos cambios para aclarar, por un lado, la asignación de responsabilidades y funciones entre *proveedores, usuarios y otros agentes de la cadena de valor de la IA*, dado que los sistemas de IA se desarrollan y distribuyen mediante cadenas de valor complejas, y por otro lado, los supuestos en que otros agentes de la cadena de valor debían asumir las responsabilidades del proveedor⁶⁷³.

Tales recomendaciones han sido finalmente incluidas en el texto definitivo del RIA, tal como se explica en los Considerandos 83 y siguientes. Baste señalar, por el momento, que en el RIA se hace especial hincapié en que las numerosas partes que intervienen en la cadena de valor suministrando sistemas, herramientas, servicios, componentes o procesos que el proveedor incorpora al sistema de IA con diversos objetivos (como el entrenamiento de modelos, el reentrenamiento de modelos, la prueba y evaluación de modelos, la integración en el *software* u otros aspectos del desarrollo de modelos) *deben proporcionar a dicho proveedor, mediante acuerdo escrito, la información, las capacidades, el acceso técnico y demás asistencia que sean necesarios habida cuenta del estado actual de la técnica generalmente reconocido, a fin de que el proveedor pueda cumplir íntegramente las obligaciones establecidas en el presente Reglamento, pero sin comprometer, no obstante, sus propios derechos de propiedad intelectual e industrial o secretos comerciales*. A nuestro modo de ver, la obligación de que los distintos operadores de la cadena de valor deban facilitarse esta información

⁶⁷² En el Considerando 17 se establece que «*El gran número de personas que suele participar en el diseño, el desarrollo, la introducción generalizada y el funcionamiento de sistemas de IA de alto riesgo hace difícil que los perjudicados identifiquen a la persona potencialmente responsable de los daños causados y demuestren que se cumplen las condiciones para interponer una demanda por daños y perjuicios*».

⁶⁷³ Con este fin, se añaden algunas disposiciones adicionales en los arts.13 y 14 que permiten una cooperación más eficaz entre los proveedores y los usuarios. Declara el Consejo que «El texto transaccional también pretende aclarar la relación entre las obligaciones establecidas en virtud del Reglamento de Inteligencia Artificial y las obligaciones ya vigentes en virtud de otros actos legislativos, como la legislación pertinente de la Unión en materia de protección de datos o la legislación sectorial, también en lo que respecta al sector de los servicios financieros. Además, el nuevo artículo 23 bis define con mayor claridad los supuestos en los que otros agentes de la cadena de valor están obligados a asumir las responsabilidades del proveedor». Vid., Orientación General del Consejo en <https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/es/pdf>, pp. 5-6.

jugará un papel importante a la hora de determinar quién de ellos puede ser el responsable del resultado final del sistema de IA.

Dando un paso más, el hecho de que estos sistemas no puedan considerarse productos terminados tras su comercialización al estar en constante evolución y sujetos a mejoras o actualizaciones para su correcto funcionamiento, conlleva que los posibles responsables puedan ir variando desde su introducción en el mercado o puesta en servicio, apareciendo nuevos sujetos responsables a lo largo de su ciclo de vida tanto por alteraciones introducidas en él, como por el no seguimiento de las instrucciones de mantenimiento o actualizaciones.

A mayor abundamiento, la atribución de responsabilidad se complica cuando se utilizan sistemas con un grado alto de autonomía decisional, ya que el humano puede llegar a perder totalmente la dirección y control sobre las decisiones adoptadas por el mismo, en cuyo caso no es sencillo imputar la responsabilidad a una persona física o jurídica concreta, al haber sido el sistema (y no el humano) el que ha generado la «salida» sin seguir instrucciones previas⁶⁷⁴. Teniendo en cuenta que en todos los ordenamientos jurídicos de la Unión el centro último de imputación de la responsabilidad es una conducta o actividad *humana*, el hecho de no poder imputar el daño al humano (al proceder la acción u omisión culposa de la «máquina») podría conllevar que el perjudicado por el *machine learning* quedara sin resarcir⁶⁷⁵.

No obstante, sobre esta cuestión debe partirse de una premisa importante sentada por el Parlamento Europeo en su *Resolución sobre un régimen de responsabilidad civil en materia de inteligencia artificial de 20 de octubre de 2020*, con la que se aleja de la previamente mantenida en su *Resolución sobre normas de derecho civil sobre robótica de 16 de febrero de 2017* y se aproxima a la postura defendida por la Comisión tanto en el *Informe del Grupo de Expertos en responsabilidad y nuevas tecnologías, de 21 noviembre 2019*, como en el que acompañaba al Libro Blanco sobre la IA: el humano no puede escudarse tras la máquina o sistema inteli-

⁶⁷⁴ Esta situación ya fue puesta de relieve en los inicios de los trabajos europeos por el Grupo de Trabajo del PE en su Proyecto de Informe con Recomendaciones a la Comisión (2016) en cuyo Considerando S se advertía de que cuánto más autónomos fuesen los robots, menos se los podía considerar simples instrumentos en manos de otros agentes (como el fabricante, el propietario, el usuario, etc.) y por ello la normativa general sobre responsabilidad resultaba insuficiente y precisaba de nuevas normas centradas en cómo una máquina podía considerarse parcial o totalmente responsable de sus actos u omisiones.

⁶⁷⁵ Un estudio de Derecho comparado sobre esta cuestión puede verse en KARNER, E./ KOCH, B. /GEISTFELD, M., *Comparative Law Study on Civil Liability for Artificial Intelligence*, estudio encargado por la Comisión Europea, Dirección General de Justicia y Consumidores, al *European Law Institute (ELI)*, Publications Office, Publications Office, 2021, disponible en: <https://data.europa.eu/doi/10.2838/66412>

gente para exonerarse de responsabilidad, ya que, «*todas las actividades, dispositivos o procesos físicos o virtuales gobernados por sistemas de IA pueden ser técnicamente la causa directa o indirecta de un daño o un perjuicio, pero casi siempre son el resultado de que alguien ha construido o desplegado los sistemas o interferido en ellos*»⁶⁷⁶.

Tres años después de su inicial propuesta de atribuir personalidad jurídica a los sistemas de IA para poder imputarles responsabilidad, la considera obsoleta e innecesaria⁶⁷⁷, y mantiene que, aun cuando se trate de sistemas de aprendizaje profundo, el operador no podría exonerarse de responsabilidad alegando que carecía de control sobre el sistema y que fue causado por un funcionamiento autónomo del mismo⁶⁷⁸. Se zanja así, por el momento, el debate acerca de la conveniencia de otorgar personalidad jurídica a los sistemas de IA a fin de que pudieran asumir obligaciones, postura que fue objeto de numerosas críticas⁶⁷⁹.

⁶⁷⁶ Observación n° 7 de la Resolución (ap. *Responsabilidad civil e inteligencia artificial*).

⁶⁷⁷ En el Considerando 6 de la Propuesta de Reglamento que acompaña a la Resolución, se justifica la innecesariedad de atribuirles personalidad jurídica alegando que muchos sistemas de IA tampoco son tan diferentes de otras tecnologías basadas en programas informáticos aún más complejos, y que en última instancia la gran mayoría de los sistemas de IA se utiliza para efectuar tareas triviales, sin riesgo o con riesgos mínimos para la sociedad.

⁶⁷⁸ Esta misma idea la reitera el CESE en su *Dictamen sobre la propuesta de Directiva sobre responsabilidad en materia de IA*, al señalar que la responsabilidad del diseño inicial y la responsabilidad última por sus posibles fallos deben seguir recayendo en los agentes humanos, tal como reconoce el Grupo de Expertos de Alto Nivel sobre IA en sus Directrices éticas. También la AEPD se pronuncia sobre el tema en su informe sobre *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial...*, cit., p. p. 19, donde afirma que no es aceptable en ningún caso trasladar la responsabilidad al propio sistema de IA. Y la UNESCO en su *Recomendación sobre la ética de la IA*, asimismo declara que hay que partir de la base de que «un sistema inteligente nunca puede reemplazar la responsabilidad final de los seres humanos y su obligación de rendir cuentas, y debería evitarse que las decisiones de vida o muerte sean cedidas a sistemas inteligentes».

⁶⁷⁹ Téngase en cuenta que la Carta abierta a la Comisión Europea sobre IA, firmada por 285 expertos en la materia de 14 Estados miembros, se manifestó totalmente en contra a la creación de cualquier tipo de personalidad jurídica para los robots. Se afirma que la creación de esta categoría se basa en una sobreestimación de las aptitudes reales de los sistemas de IA y que no hay ninguna categoría jurídica actual a la que asemejar la nueva. La carta se puede consultar en: <http://www.robotics-openletter.eu>. En contra de esta personificación jurídica de los robots y sistemas de IA, por todos, ANGUIA RÍOS, R., «Inteligencia artificial y Derecho civil: líneas de pensamiento en materia de daños», *Revista Crítica de Derecho Inmobiliario*, n° 781, 2020, pp. 2559 y ss., y p. 2567 o SUÑÉ LLINAS, E. *Derecho e Inteligencia Artificial: de la robótica a lo posthumano*, Valencia, 2020, p. 117, para quien, dadas las características actuales de los sistemas de IA, dotarlos de personalidad jurídica sería simplemente un medio para poder imputarles la responsabilidad civil, liberando de responsabilidad a las grandes corporaciones de IA y la robótica de sus propia responsabilidad a través del nuevo responsable ficticio. No obstante, no niega que a largo plazo puede contemplarse algún tipo de personalidad jurídica específica para los robots.

Esto va en la línea de lo que se mantenía ya en el Informe NTF 2019 que establecía como regla que, en virtud del *principio de equivalencia funcional*, un usuario de una tecnología que tiene un cierto grado de autonomía no debería ser menos responsable de los daños resultantes que si dichos daños hubieran sido causados por un auxiliar humano. En nuestro ordenamiento, esta previsión equivaldría a la responsabilidad por hecho ajeno contemplada del art. 1903.4 Cc, esto es, la responsabilidad de los dueños o directores de establecimiento o empresa respecto de los perjuicios causados por sus dependientes. Este principio, junto con el de *neutralidad tecnológica*, en virtud del cual las reglas propias de un ordenamiento jurídico no se deben modificar en función de la tecnología sobre la que han de proyectarse, van dirigidos a lograr que las víctimas de daños causados por la IA obtengan una protección equivalente a las víctimas de daños causados por los demás productos o en los casos en que no medien sistemas de IA.

Tales principios vuelven a ocupar un lugar protagonista en las últimas Propuestas de Directivas sobre responsabilidad de la Comisión, de 28 de septiembre de 2022. Se trata de garantizar, a fin de cuentas, el derecho a la tutela judicial efectiva consagrado en el art. 47 de la CEDF.

A nuestro modo de ver, esta postura es acertada en la medida en que, al menos en la realidad actual, siempre se precisa en mayor o menor medida la intervención humana en las diferentes fases de creación, desarrollo, implementación y uso del modelo⁶⁸⁰. Además, el nuevo RIA no parece admitir la autonomía plena de los sistemas ya que obliga a adoptar medidas de vigilancia efectiva por personas físicas durante el período de su utilización. Así, dispone su art. 14 que los sistemas de alto riesgo se diseñarán y desarrollarán de tal manera que sean supervisados eficazmente por personas físicas en proporción a sus riesgos, nivel de autonomía y contexto de uso.

⁶⁸⁰ Manteniendo que los sistemas de IA siempre precisan intervención humana para aprender y entrenar, entre otros, ORTIZ FERNÁNDEZ, M., *La responsabilidad civil derivada de los daños causados por sistemas inteligentes y su aseguramiento: análisis del tratamiento ofrecido por la Unión Europea*, Madrid 2021, p. 98 o LLORENS ESPADA, J., *op.cit.*, p. 590, quien afirma que «incluso cuando los sistemas de inteligencia artificial escapan del control de las decisiones humanas, sus decisiones provienen de una previa delegación humana, lo cual permite encontrar una persona como responsable último de la decisión». Sigue diciendo este autor, citando a VALVERDE ASENSIO, que «El carácter intransmisible del poder de dirección genera que la empresa no pueda alegar ningún mecanismo de exculpación de su responsabilidad como consecuencia de la delegación de decisión en un algoritmo». Para ANGUITA RÍOS, R., *op.cit.*, p. 2566, aunque debe distinguirse entre la actividad del robot (en nuestro caso sistema de IA) que procede directamente de lo que el programador introduce, de aquello de lo que es posteriormente capaz de realizar por sí mismo, en ambos casos hay que tener en cuenta y valorar que este desarrollo posterior está relacionado con el aprendizaje inicial y la actividad humana que lo facilitó.

4.2. CONSIDERACIONES A TENER EN CUENTA SOBRE LA ATRIBUCIÓN DE RESPONSABILIDAD

En base a todo lo afirmado, se pueden sentar dos máximas a tener en cuenta en este punto:

- A) *A mayor grado de autonomía del sistema, mayor complejidad para atribuir responsabilidad al humano*

Las dificultades para determinar el sujeto responsable de los daños ocasionados por las decisiones algorítmicas aumentan en paralelo a la autonomía y capacidad de autoaprendizaje del sistema. En consecuencia, los sistemas de toma de decisiones basados en algoritmos deterministas no implicarán grandes problemas para determinar quién es el ser humano responsable, al estar predeterminados sus resultados en base a unas instrucciones previas del programador⁶⁸¹; por el contrario, en el caso de los sistemas autónomos que tienen margen de decisión y son dinámicos o adaptativos, resultaría mucho más difícil atribuir la responsabilidad a un ser humano en concreto.

Por lo que se refiere a estos últimos, se han propuesto por los autores varias teorías y argumentos para solucionar el tema de la imputación de la responsabilidad⁶⁸². Entre ellas, aplicar por analogía las reglas de la responsabilidad civil por daños causados por animales con base en el art. 1905 Cc, las de daños causados por cosas peligrosas de los arts. 1908 o 1910 Cc, o las de la responsabilidad por hecho ajeno del art. 1903 Cc.

Respecto a esta última, se ha defendido concretamente la aplicación del art. 1903.4 Cc, esto es, la responsabilidad del empresario por los hechos de sus dependientes que, como se sabe, es un supuesto excepcional en que se debe responder por hechos ajenos. Como explica NAVAS NAVARRO, cuando el sistema experto actúa como «auxiliar» del humano, entonces parece que la responsabilidad de este último se asemeja a la responsabilidad «vicaria» del principal por los hechos de sus auxiliares. En cambio, cuando actúa de forma autónoma tomando sus propias decisiones, a partir de la actuación del humano pero sin seguir sus instrucciones, la similitud con la responsabilidad del principal por

⁶⁸¹ Señala NAVAS NAVARRO, S., «Sistemas expertos ...», *cit.*, p. 3/16, que en los casos de daños por sistemas que siguen un patrón determinista (sin capacidad de decisión ni autonomía) cuya lógica computacional no plantea problemas en la medida en que se trata de una lógica formal «si... entonces», la responsabilidad civil sería por hecho propio y no por hecho ajeno y debería mantenerse la regla general de responsabilidad por culpa (art. 1902 CC).

⁶⁸² Puede verse una explicación detallada de cada una de ellas en ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...* *cit.*, capítulos III y IV.

los hechos de sus auxiliares desaparece. No obstante, en su opinión, aun siendo esto así debería mantenerse este fundamento de la responsabilidad⁶⁸³. Propone atribuir a los sistemas de IA la condición de «sujeto de derecho», lo que, según aclara, no implicaría atribuirles previa y necesariamente personalidad jurídica. Con ello, se podría establecer el juicio de imputabilidad con el sistema de IA mientras que el juicio de responsabilidad se establecería con el humano, el cual podría ser propietario o poseedor del sistema de IA (especialmente, el usuario del mismo). Por tanto, en su opinión, atribuir dicha condición a los sistemas de IA solo tendría virtualidad para hacerles «imputables» de conductas respecto de las cuales «respondería» un humano: *«En este caso, la responsabilidad por hecho ajeno sería el fundamento de responsabilidad adecuado pues debemos tener en cuenta la actuación de dos posibles sujetos de derecho, el humano y el sistema de IA, de suerte que la actividad o conducta de cada uno de ellos es relevante para establecer si existe responsabilidad o no»*. Incluso apunta esta autora que, aun cuando no se dieran los presupuestos de aplicación de la responsabilidad del principal por los hechos de sus auxiliares, y en concreto, la subordinación o dependencia a las instrucciones del principal en tanto en cuanto el sistema IA hubiera actuado de forma autónoma en virtud de su capacidad de autoaprendizaje y sin seguir instrucciones previas, podría seguir haciéndose responsable al empresario (principal) por cuanto es él quien disfruta del beneficio que genera el sistema de IA⁶⁸⁴.

B) *A mayor grado de automatización, ¿mayores problemas para determinar el sujeto responsable?:*

La respuesta a esta pregunta va a depender de si los sistemas son semiautomatizados o plenamente automatizados. En otras palabras, de si sus resultados se utilizan como mero apoyo a la toma de decisión humana o si, por el contrario, sus resultados se adoptan y se ejecutan de forma totalmente automática, es decir, sin intervención humana.

Si a pesar de haber empleado un sistema de IA con un alto grado de autonomía decisional y complejidad, éste se utiliza simplemente como apoyo a la decisión, parece lógico que la responsabilidad por los daños que deriven de una decisión discriminatoria sería del sujeto o sujetos que tomaron la decisión final, tanto si ésta coincide con el resultado que arrojó el algoritmo, como si

⁶⁸³ Vid., NAVAS NAVARRO, S., «Sistemas expertos ...», *cit.*, p. 5/16.

⁶⁸⁴ Explica todo esto en *Daños ocasionados por sistemas de inteligencia artificial: especial atención a su futura regulación*, Granada, 2022, pp. 56 y ss. En particular, pueden verse las pp. 60-61. Concluye en la p. 62, que esta asunción por parte del empresario de las consecuencias dañosas del sistema IA supondría un incentivo para desplegar un mayor control de los riesgos que pueden resultar de la actividad de aquél en la medida en que se encuentra en mejor posición para establecer mecanismos de prevención (por ej. auditorías, modificaciones del código fuente, etc) y hacer frente a los costes tanto de la prevención como de la indemnización.

no. En estos casos, se trataría de una responsabilidad por hecho propio del encargado de tomar la decisión, aun cuando la misma estuviera basada en la previa decisión o recomendación (sesgada) del sistema.

Estos supuestos quedan expresamente excluidos del ámbito de aplicación de la proyectada Directiva sobre responsabilidad en materia de IA, en cuyo Considerando 15 se explica que «[n]o es necesario abarcar las demandas de responsabilidad cuando los daños hayan sido causados por una evaluación humana seguida de una acción u omisión humana y el sistema de IA se haya limitado a proporcionar información o asesoramiento que fue tenido en cuenta por el agente humano de que se trate». En tales casos, sigue señalando la Comisión, «es posible atribuir los daños causados a una acción u omisión humana, ya que la información de salida del sistema de IA no se interpone entre la acción u omisión humana y el daño, por lo que establecer la causalidad no es más difícil que en situaciones en las que no interviene un sistema de IA». En definitiva, el comportamiento posterior del sujeto que adopta la decisión final interrumpiría el nexo causal o podría considerarse la causa adecuada del daño⁶⁸⁵.

Ahora bien, si el encargado de tomar la decisión (por ejemplo, el banco que debe conceder o no el crédito) se hubiera apoyado en los resultados de un sistema que, por un defecto intrínseco (ej. mal diseño, datos sesgados, etc.), arrojará decisiones discriminatorias, habría que plantearse la posible reclamación de responsabilidad al proveedor, desarrollador, u otro agente de la cadena de valor, ante lo que podría considerarse una falta de conformidad del producto o unos vicios ocultos.

A nuestro modo de ver, lo que habría que dilucidar a estos efectos sería el tipo de contrato celebrado entre los operadores, en particular, entre el proveedor y el responsable del despliegue (o incluso entre el proveedor y el desarrollador del sistema de IA si no son la misma persona –posibilidad que contempla el art. 3.3 RIA–) a fin de determinar su régimen jurídico, ya que, en función de si, por ejemplo, se hubiera desarrollado un *software* individualizado siguiendo las instrucciones del responsable del despliegue que posteriormente lo implementa en su empresa (instrucciones que podrían ser discriminatorias⁶⁸⁶), podríamos estar ante un contrato de encargo o creación de un

⁶⁸⁵ Así lo entiende RUBÍ PUIG, A., «Responsabilidad civil e Inteligencia artificial. Un examen crítico...», *cit.*, p. 274, a cuya postura nos adherimos.

⁶⁸⁶ Pensemos que el empresario puede encargar, por ejemplo, que le diseñen un sistema de selección y cribado de currículums donde una de las instrucciones que se le dé al algoritmo sea descartar los currículums de personas de cierto sexo o raza. En estos casos, el algoritmo sería un mero ejecutor de la *orden de discriminar* del empresario. Esto fue lo que ocurrió, a nuestro modo de ver, en el caso de la empresa Deliveroo y su algoritmo Frank. El algoritmo simplemente hacía lo que se le ordenó hacer, es decir, ejecutaba el sistema de puntuación establecido por la empresa en las condiciones de trabajo de los repartidores. Para CASTILLO PARRILLA, J.A., *op.cit.*, esto «nos lleva a situar

programa informático *ad hoc*; mientras que si se hubiera adquirido un *software* ya creado y entrenado con una finalidad concreta (es decir, un programa *standard*), se trataría a nuestro juicio de un contrato de compraventa⁶⁸⁷. En este último caso habría que diferenciar, a su vez, si el desarrollador del *software*, una vez vendido el mismo, se desvincula de su mantenimiento y/o actualización o

el pleito en una cuestión discriminatoria “de toda la vida”, ejecutada a través de un algoritmo. Como puede verse, se trata de un caso de discriminación a trabajadores a través de un algoritmo, no de una discriminación del algoritmo». Considera RIVAS VALLEJO, P., «Herramientas desde el derecho antidiscriminatorio...», *cit.*, p. 562, que cuando los programadores no se hubieran limitado a trasladar a lenguaje técnico las órdenes empresariales, habría que valorar su posible responsabilidad por las decisiones del sistema, y en los casos en que conscientemente hubieran creado el algoritmo con parámetros claramente sesgados (por ejemplo, con indicación explícita de excluir a mujeres para un puesto de trabajo), también se plantearía dilemas éticos.

⁶⁸⁷ NAVAS NAVARRO, S., *Daños ocasionados por sistemas de inteligencia artificial...*, *cit.*, p. 81, diferencia entre «software predeterminado» (contenido digital) y el personalizado o elaborado para cubrir necesidades específicas del cliente («servicio digital»). En los casos en que se utilice tecnología programada y entrenada por una tercera empresa para finalidades concretas (lo que en inglés se conoce como *off-the-self algorithms* o *Artificial intelligence as a service*), considera GINÈS i FABRELLAS, A., *op.cit.*, p.320, que surge un problema adicional de responsabilidad. Según la autora, en estos casos, que suelen ser frecuentes en la práctica porque no muchas empresas disponen de los recursos y conocimientos de las grandes corporaciones tecnológicas para crear, entrenar y testear sus propios algoritmos, surge la duda de *a quién podría imputarse responsabilidad por el uso de un sistema algorítmico sesgado que genera efectos discriminatorios: si a la empresa que adopta la decisión discriminatoria o a la empresa que lo ha diseñado, creado y entrenado*. En su opinión, que coincide con la expuesta por nosotros en el texto, la empresa que utiliza el sistema resulta responsable por la decisión adoptada sin perjuicio de la acción civil que, en su caso, le correspondiera después. Es decir, sin perjuicio de que el algoritmo hubiera sido diseñado y programado por una tercera empresa, el ámbito de responsabilidad de la empresa continúa siendo las decisiones sobre gestión de las personas por ella contratadas, independientemente que dicha decisión se tome por medios humanos o automatizados. Para esta autora, «el uso de la IA está incrementando la dirección algorítmica o el “Black Box Boss” lo que requiere que las personas encargadas de utilizar y aplicar dicha tecnología tengan los conocimientos técnicos suficientes para entenderla –cuanto menos, entender la lógica de funcionamiento– y para apreciar su potencial impacto discriminatorio y la responsabilidad que puede derivarse». Por ello, considera que las personas que ocupan cargos de dirección deben tener conocimientos tecnológicos suficientes para entender el funcionamiento de los sistemas de IA que implementan en su empresa y poder evitar situaciones discriminatorias. Para RIVAS VALLEJO, P., «Herramientas desde el derecho antidiscriminatorio...», *cit.*, p. 543, en los casos de propiedad del *software* ajena, podría realizarse alguna propuesta de *lege ferenda* con el fin de modular el grado de responsabilidad por el uso de un *software* sesgado en función de la tipología los sesgos que presente y, por ende, del conocimiento empresarial sobre los sesgos indirectos (ya sea por error, por derivaciones del aprendizaje profundo, etc.) o su impacto discriminatorio (lo que exigiría imponer la obligación de realizar análisis de impacto de riesgo algorítmico previo a su implantación en la empresa y auditorías periódicas, en la línea de la prevención de riesgos para la salud prevista en la Ley 31/1995, de Prevención de Riesgos Laborales. Y según ANGUITA RIOS, R.M., *op.cit.*, p. 2567, ante un daño causado por un robot (en nuestro caso, un *software* que no está incorporado a un bien corpóreo [*stand-alone-software*]), la responsabilidad objetiva devendrá primeramente para el dueño del robot y, una vez resuelto este primer elemento, se realizará la depuración de responsabilidades en función de las razones que operen tras el actuar dañoso.

si se obliga a ello (en cuyo caso estaríamos ante un *contrato complejo de compraventa y prestación de servicios*, según ha señalado en alguna ocasión el TS)⁶⁸⁸.

También podría ocurrir, como es frecuente, que la relación que une al proveedor y al responsable del despliegue se enmarcara en el ámbito de la mera prestación de servicios. Así ocurre, por ejemplo, cuando las empresas de soluciones de información y análisis de riesgo (como por ejemplo la agencia SCHUFA del caso resuelto por la STJUE de 7 de diciembre de 2023 anteriormente comentada) analizan, a petición de otras empresas (por ejemplo, un banco), la solvencia crediticia de un posible prestatario a partir de la información de la que disponen sobre él y/o de la información que sus socios comerciales (en nuestro ejemplo, el banco) le proporcionan. El banco contrata ese servicio con la agencia de crédito a fin de que ésta le proporcione un informe de solvencia del prestatario, con ayuda del cual decidirá la celebración o no del contrato, o fijará las condiciones en las que el mismo se realizará (en nuestro ejemplo, el informe puede servir al banco no sólo para decidir si concede o no el préstamo, sino también bajo qué condiciones lo concede –tipo de interés, tiempo, comisiones de amortización, etc–.). Si los informes de solvencia no son suficientemente precisos (o están sesgados en detrimento de cierto grupo social), el banco que basándose en ellos deniega un crédito por motivos discriminatorios debería responder ante el potencial prestatario por los daños ocasionados (responsabilidad que sería extracontractual o precontractual), pero, bajo nuestro punto de vista, después podría reclamar responsabilidad a la agencia de crédito por no cumplir correctamente con su prestación contractual (responsabilidad contractual), más si cabe cuando los informes proporcionados fueran determinantes para la toma de decisión –como de hecho lo son en la mayoría de ocasiones–. Si las entidades de información y análisis de riesgos hubieran adquirido, a su vez, el *software* defectuoso a otro proveedor tecnológico, podrían accionar contra éste.

⁶⁸⁸ Según señala PLAZA PENADES, J., «Los contratos informáticos y electrónicos», en J.R. De Verda y Beamonte (coord.) *Derecho civil II. Obligaciones y contratos*, Valencia, 2021, p. 570, en los contratos de encargo o creación de un programa de ordenador individualizado o personalizado existe, salvo pacto en contrario, la obligación de revelar el código fuente al que lo ha encargado. En este sentido, cita la STS 17 de 17 de mayo de 2003 (Tol 274.503) que se pronuncia sobre un contrato de instalación integral de un sistema informático para dos centros de trabajo que incluía el suministro de un programa informático individualizado y el mantenimiento (al que considera un complejo contrato de compraventa y arrendamiento de servicios) y declara que en este supuesto «habida cuenta de que el programa que se convino, fue uno creado por la suministradora “ad hoc”, para que cubriese las necesidades informáticas del cliente [...] y que [...] por tratarse de un programa personalizado, y haber corrido el cliente con todos los gastos de investigación y desarrollo del programa, la proveedora debió de entregarle una copia de las fuentes, ya que sin ella no se puede actualizar el programa hecho a medida, ni introducir posibles mejoras».

Hechas las anteriores consideraciones, habría que preguntarse ahora si podría llegarse a la misma conclusión en los casos en que el responsable del despliegue, en lugar de apoyarse en la puntuación o valor de probabilidad derivado del análisis de riesgo crediticio realizado por el sistema de IA, delegara totalmente la decisión de conceder o denegar el crédito en dicho sistema. En este caso, siempre que se contara con base legitimadora para ello en virtud del art. 22 RGPD⁶⁸⁹, el sistema no sólo perfilaría al potencial prestatario en un nivel de riesgo concreto, sino que junto con eso le concedería o denegaría automáticamente el crédito, sin que hubiera una *intervención significativa* de un trabajador del banco que revisara esa decisión, o que la valorara junto con otros factores para tomar la decisión final. A nuestro juicio, en ambas hipótesis el responsable del despliegue respondería de la decisión plena o semiautomatizada del sistema frente al cliente-consumidor al que se hubiera denegado el crédito –o concedido en peores condiciones– por motivos discriminatorios. La razón estriba, principalmente, en que en ambos casos el responsable del despliegue es el que acaba discriminando, es decir, es la decisión del empresario lo que ocasiona el daño; pero además, es el que tiene el control sobre la implementación del sistema en su empresa, es decir, puede decidir si lo utiliza o no y cómo lo utiliza (como apoyo a la decisión o delegando en él plenamente la decisión en los casos legalmente permitidos); si decide usarlo por el motivo que sea –por ejemplo, para mejorar la efectividad o precisión de sus decisiones, para ahorrarse tiempo y recursos en la realización de tareas, etc.– a nuestro modo de ver debería asumir los riesgos que genere dicho uso frente a los perjudicados que, en este caso, se traducirían en resultados ilícitos, sesgados o incorrectos que perjudican a la persona.

Cabría preguntarse también si, frente a la posterior reclamación de responsabilidad dirigida por el responsable del despliegue al proveedor de la tecnología, éste podría alegar que el primero era conocedor de los riesgos que implicaba el uso del sistema y aun así decidió implementarlo en su empresa⁶⁹⁰. ¿Podría considerarse negligente usar este tipo de sistemas altamente precisos aun a sabiendas de que pueden contener sesgos implícitos o inferidos que lleven a resultados discriminatorios? Un sector doctrinal considera que, partiendo del deber de los empresarios de prevenir riesgos laborales y de la obligación de no discriminar, sería negligente emplear estos sistemas para tomar decisiones con efectos directos sobre las personas y deberían responder

⁶⁸⁹ Si no fuera así, se estaría infringiendo también el RGPD con las consecuencias que ello implica en cuanto a la responsabilidad por daños del art. 82 RGPD y las posibles sanciones pecuniarias.

⁶⁹⁰ Téngase en cuenta que, como ya analizamos en otro momento, en virtud del art. 13 RIA existen ciertas obligaciones de transparencia e información de comunicación de los proveedores de sistemas de IA de alto riesgo hacia los responsables del despliegue con el fin de que puedan interpretar y usar correctamente su información de salida. Vid., concretamente el art. 13.3.b)

de los daños que se generen⁶⁹¹. No obstante, para otro sector doctrinal, la utilización de estos sistemas no podría considerarse una negligencia en sí misma, y ni siquiera una conducta peligrosa o arriesgada, partiendo de que se trata de sistemas o tecnología legalmente comercializados y que han superado los estándares de seguridad pertinentes para su puesta en el mercado⁶⁹².

Por último, habría que plantearse si sería posible que el perjudicado por la decisión discriminatoria del sistema reclamara responsabilidad contra otros sujetos como el fabricante del mismo, al amparo de la normativa sobre productos defectuosos⁶⁹³. Si bien creemos que en los casos que venimos considerando lo más normal en la práctica será dirigir la acción directamente contra la empresa o Administración pública que adoptó la decisión discriminatoria que perjudica a la persona (v.gr., banco, hospital, aseguradora, etc.), no habría que descartar una posible reclamación por esta vía siempre y cuando se dieran los presupuestos para ello, principalmente el defecto intrínseco del sistema que en nuestro caso podría ser, por ejemplo, un mal diseño que generara sistemáticamente resultados discriminatorios. Ahora bien, también creemos que articular la reclamación por esta vía no tendría utilidad en la generalidad de los casos en que se reclamen daños derivados de la discriminación algorítmica teniendo en cuenta que la normativa de productos defectuosos no cubre los daños prototípicos que se generan en dichos supuestos, esto es, los daños morales y patrimoniales puros⁶⁹⁴. En cualquier caso, habrá que estar al caso concreto para elegir la vía más adecuada para articular la reclamación.

⁶⁹¹ SÁEZ LARA, C., *op.cit.*, pp. 185-186 y p. 196.

⁶⁹² En este sentido ATAZ LÓPEZ, J., «Daños causados por las cosas...», *cit.*, pp. 32-33. ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...* *cit.*, trata con detalle este tema en las pp. 206 y ss. Para esta autora, la utilización correcta de sistemas de IA autónomos y con capacidad de autoaprendizaje no revela en sí misma negligencia alguna. Admitidos los beneficios que aportan tales sistemas a la sociedad y comprobado que no incrementan el riesgo de producir daños en los sectores en que se introducen, su mera creación o posterior utilización no supone una negligencia de sus creadores o usuarios por el mero hecho de que estos sistemas hayan actuado con autonomía causando daños.

⁶⁹³ La PDPD señalaba en su Considerando 12 que el desarrollador o productor de programas informáticos, incluidos los proveedores de sistemas de IA, debían ser tratados como fabricantes.

⁶⁹⁴ Sobre esta cuestión nos remitimos al epígrafe 2 del Capítulo V.

Capítulo IV

Medidas para hacer frente a los problemas que plantea la discriminación algorítmica (I): la prevención del daño

1. INTRODUCCIÓN: ABORDANDO LOS RIESGOS DE LA IA DE FORMA INTEGRAL E INTERDISCIPLINAR

Como se ha ido viendo en los apartados anteriores, tanto la opacidad de los sistemas de IA como el resto de sus características intrínsecas, dificultan la averiguación de *cuándo, cómo y por qué* se ha originado la discriminación algorítmica y *quién* debería responder de los daños que cause. La complejidad para determinar tales cuestiones una vez producido el daño puede conllevar incluso que las víctimas queden sin resarcir. Por ello, deviene necesario adoptar medidas dirigidas a prevenirlos mediante el control *ex ante* de los sistemas de IA.

En esta línea se dirigen la mayoría de propuestas normativas nacionales, europeas e internacionales en la materia. Principalmente, este enfoque preventivo del daño fue el que ya en 2021 acogió la Comisión Europea en su inicial propuesta de RIA, donde instauraba una regulación basada en el riesgo que podían implicar los sistemas de IA para la salud, seguridad o derechos fundamentales de las personas, estableciendo *normas dirigidas a minimizar el riesgo de que tales sistemas causasen daños*⁶⁹⁵. También en plena consonancia con el RIA se sitúa el *Convenio Marco del Consejo de Europa sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho*, cuyo objetivo principal es pro-

⁶⁹⁵ También se mantenía dicho enfoque de prevención de daños, por ejemplo, en el *Informe de la Comisión al Parlamento Europeo, al Consejo y al Comité Económico y Social Europeo, sobre las repercusiones en materia de seguridad y responsabilidad civil de la inteligencia artificial, el internet de las cosas y la robótica*, de 19 de febrero de 2020 (COM/2020/64 final).

porcionar un marco eficaz a nivel internacional con el que *prevenir y mitigar* los riesgos que plantea la IA para los derechos humanos, la democracia y el Estado de derecho.

No obstante, junto con las medidas preventivas, es lógico que se articulen también medidas reparadoras para el caso de que, a pesar de aquellas, el daño llegara a producirse.

No son pocos los problemas que plantea el resarcimiento de los daños causados por estos sistemas, a los que las instituciones europeas han tratado de dar respuesta en los últimos tiempos. Ya en 2012, un estudio financiado por la Comisión Europea sobre los aspectos jurídicos y éticos de la regulación de las tecnologías robóticas emergentes en Europa (conocido como *ROBOLAW* «*Regulating emerging robotic technologies in Europe: Robotics facing Law and Ethics*»), identificó el tema de la responsabilidad civil por daños como una de las cuestiones más relevantes que planteaban estas tecnologías⁶⁹⁶.

Tras dicho estudio, la principal iniciativa europea en materia de responsabilidad civil en el ámbito de la IA se recogió en la *Resolución del Parlamento Europeo, de 20 de octubre de 2020* (2020/2014(INL)) a la que se anexaba una propuesta de Reglamento que regulaba el tema⁶⁹⁷, sin perjuicio de la existencia del régimen de responsabilidad aplicable a los productos defectuosos recogido en la Directiva 85/374/CEE, cuya adaptación a las exigencias y particularidades de la IA se consideraba imprescindible.

No obstante, el 28 de septiembre de 2022 la Comisión Europea lanzó dos propuestas regulatorias sobre responsabilidad civil: por un lado, la *Propuesta de Directiva para adaptar las normas de responsabilidad extracontractual en la Inteligencia Artificial (AI Liability Directive)*, y por otro lado, la *Propuesta de Directiva para modernizar el régimen de responsabilidad por productos defectuosos*. Su objetivo era promover la confianza en la IA garantizando que, si a pesar de los requisitos preventivos impuestos en el RIA y otras normas de seguridad se producían daños, las víctimas podrían recibir una indemnización efectiva. Se parte de que sólo mediante esta doble protección, tanto *ex ante* como *ex post*, se podrá conseguir que los ciudadanos confíen en esta tecnología. Confianza que, además, debería recaer tanto en la propia tecnología (asegurando una baja probabilidad de fallo), como en el ordenamiento jurídico (garantizando medidas para evitar daños y posibilidad de reparación en caso de sufrirlo).

⁶⁹⁶ Los objetivos de este proyecto eran dos: por un lado, identificar las principales cuestiones jurídicas y éticas que planteaban las tecnologías robóticas emergentes y valorar si el marco existente para hacerles frente era adecuado; y por otro lado, aclarar cómo podían afectar los avances en robótica a los valores sociales imperantes en Europa.

⁶⁹⁷ Propuesta de Reglamento del Parlamento Europeo y del Consejo relativo a la responsabilidad civil por el funcionamiento de los sistemas de inteligencia artificial.

Así pues, con el RIA y las Directivas sobre responsabilidad civil se pretendía abordar de forma *integral* los riesgos que plantea la IA, estableciendo disposiciones dirigidas tanto a prevenirlos como a corregirlos, respectivamente. Tal como afirma la Comisión en la Exposición de Motivos de su PDRC: «[l]a seguridad y la responsabilidad son las dos caras de la misma moneda: intervienen en momentos diferentes y se refuerzan mutuamente. Si bien las normas para garantizar la seguridad y proteger los derechos fundamentales reducirán los riesgos, no los eliminan por completo. En caso de que se materialice tal riesgo, sigue existiendo la posibilidad de que se produzcan daños. En tales casos, se aplicarán las normas sobre responsabilidad civil de la presente propuesta. Unas normas eficaces en materia de responsabilidad también ofrecen un incentivo económico para cumplir las normas de seguridad y, por lo tanto, contribuyen a evitar que se produzcan daños. Además, la presente propuesta contribuye al cumplimiento de los requisitos para sistemas de IA de alto riesgo impuestos por la Ley de IA, ya que el incumplimiento de dichos requisitos constituye un elemento importante de los que dan lugar al aligeramiento de la carga de la prueba».

Además, estas propuestas no sólo se complementaban entre ellas, sino también con otras políticas de IA de la Comisión basadas en requisitos preventivos normativos y de supervisión destinados directamente a evitar violaciones de los derechos fundamentales, como el RGPD, la *Ley de Servicios Digitales* y la *legislación de la UE sobre no discriminación e igualdad de trato*⁶⁹⁸.

A nuestro modo de ver, las medidas europeas sobre IA vienen a cubrir algunos vacíos de protección de los que adolece la normativa de protección de datos y el Derecho antidiscriminatorio para hacer frente a la discriminación algorítmica. Por tanto, cuando se incorpore la IA al proceso de toma de decisiones basadas en el tratamiento de datos personales y las mismas resulten discriminatorias para una persona o grupo, habría que combinar las respuestas ofrecidas por la normativa antidiscriminatoria, la de protección de datos, y

⁶⁹⁸ De hecho, en el Considerando 10 RIA se dice que «El presente Reglamento no pretende afectar a la aplicación del Derecho de la Unión vigente que regula el tratamiento de datos personales, incluidas las funciones y competencias de las autoridades de supervisión independientes competentes para vigilar el cumplimiento de dichos instrumentos. Tampoco afecta a las obligaciones de los proveedores y los responsables del despliegue de sistemas de IA en su papel de responsables o encargados del tratamiento de datos derivadas del Derecho de la Unión o nacional en materia de protección de datos personales en la medida en que el diseño, el desarrollo o el uso de sistemas de IA impliquen el tratamiento de datos personales. También conviene aclarar que los interesados siguen disfrutando de todos los derechos y garantías que les confiere dicho Derecho de la Unión, incluidos los derechos relacionados con las decisiones individuales totalmente automatizadas, como la elaboración de perfiles. Unas normas armonizadas para la introducción en el mercado, la puesta en servicio y la utilización de sistemas de IA establecidas en virtud del presente Reglamento deben facilitar la aplicación efectiva y permitir el ejercicio de los derechos y otras vías de recurso de los interesados garantizados por el Derecho de la Unión en materia de protección de datos personales, así como de otros derechos fundamentales».

las previstas en el RIA y en las normativas europeas sobre responsabilidad que lleguen a adoptarse⁶⁹⁹.

En el presente capítulo y en el siguiente se analizarán, respectivamente, las medidas que se recogen en tales normativas, vigentes o proyectadas, para *prevenir* y *reparar* los daños al derecho a la igualdad de trato y no discriminación. Se expondrá también cómo se ha venido abordando a nivel europeo la regulación y el control de los riesgos que los sistemas inteligentes plantean para los derechos fundamentales y la responsabilidad por los daños que se ocasionen.

Junto con las medidas jurídicas, se prestará especial atención a otras soluciones que pueden adoptarse desde el punto de vista de la ética y desde el propio ámbito tecnológico para evitar la discriminación algorítmica. Como venimos defendiendo, solo a través de la combinación de instrumentos legales, enfoques técnicos y colaboración entre disciplinas se podrá garantizar la equidad y la no discriminación en la era de la IA.

⁶⁹⁹ Como también opina RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», *cit.*, p. 453, resulta indispensable aunar todos estos elementos en el adecuado abordaje jurídico de la cuestión. Pero hasta que tales propuestas normativas sobre IA se aprueben, entren en vigor y sean plenamente aplicables, deberá resolverse el problema de la discriminación algorítmica con el Derecho positivo a nuestro alcance, aunando la protección de datos, la tutela antidiscriminatoria y las salvaguardas frente a los obstáculos que dificultan una protección integral (por ejemplo, propiedad intelectual, secreto empresarial, opacidad mercantil...) –*vid.*, p. 543–. En esta línea, mantienen también GERARDS, J./XENIDIS, R., *op.cit.*, p. 146, que adoptar un enfoque legal multidisciplinar que integre varios campos del Derecho (derecho antidiscriminatorio, normativa de protección de datos, propiedad intelectual, protección del consumidor, etc.) y que fomente la cooperación entre varios actores legales (por ejemplo, organismos de igualdad, autoridades de protección de datos con experiencia en el campo de las nuevas tecnologías, asociaciones de consumidores, etc.) podría resultar clave para abordar adecuadamente la discriminación algorítmica. No obstante, también manifiestan sus dudas acerca de que la combinación de las respuestas ofrecidas por dichos marcos jurídicos a nivel nacional, sean suficientes para conseguir dicho objetivo.

No obstante, ha de tenerse en cuenta que no todas las soluciones de IA tratan datos personales, ni toman decisiones basadas únicamente en tratamientos automatizados que afectan a personas físicas. Por tanto, sólo si el componente IA realiza el tratamiento de datos personales, elabora perfiles sobre una persona física o si toma decisiones sobre la misma, tendrá que someterse al RGPD. En caso contrario, no será necesario. Así lo explica la AEPD en su informe sobre *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial...*, *cit.*, p. 14/53, advirtiendo de que «en muchos casos no es sencillo determinar si durante una etapa del ciclo de vida de un sistema basado en IA se tratan o no datos personales». Sobre este tema, señala RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 657, que las operaciones que traten datos personales incumpliendo la normativa en este ámbito, ya sea en la fase de desarrollo del sistema, en su funcionamiento o resultados, quedarán en la mayoría de ocasiones bajo el ámbito de aplicación del RGPD y los daños que se deriven podrían compensarse aplicando el art. 82 RGPD. Además, en el marco de una tecnología de IA específica deberá valorarse contextualmente, para poder determinar los sujetos responsables, si el proveedor o el usuario-operador de esta podrían asumir la posición de responsable o encargado del tratamiento a la hora de responder civilmente con arreglo a este régimen.

2. ANÁLISIS DE LAS PRINCIPALES NORMAS JURÍDICAS DIRIGIDAS A PREVENIR LA DISCRIMINACIÓN ALGORÍTMICA Y LOS DAÑOS QUE DE ELLA PUEDAN DERIVAR

2.1. EL REGLAMENTO DE INTELIGENCIA ARTIFICIAL (RIA): REGULANDO DESDE UN ENFOQUE PREVENTIVO

No cabe duda de que la UE ha realizado un importante esfuerzo a la hora de regular la IA mediante un enfoque garantista de los derechos fundamentales y la seguridad, que ha situado a Europa a la vanguardia del sector y en referente de otras jurisdicciones. Este prolijo e incesante trabajo de los organismos comunitarios para lograr una IA confiable, ha culminado con la aprobación del Reglamento 2024/1689, de 13 de junio de 2024⁷⁰⁰ (RIA), que vino precedida por múltiples iniciativas en forma de resoluciones, directrices, informes o recomendaciones de distintas instituciones europeas, entre las que se pueden destacar, por su importancia, la *Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones sobre normas de Derecho civil sobre robótica*, el *Informe NTF de 21 de noviembre de 2019*, el *Libro Blanco sobre la IA* y el Informe que lo acompaña, así como tres *Resoluciones del Parlamento y el Consejo, de 20 de octubre de 2020*, que incluían recomendaciones a la Comisión sobre cuestiones como la ética, la responsabilidad civil y los derechos de propiedad intelectual en el ámbito de la IA, que venían acompañadas de propuestas de Reglamentos que seguían la estela marcada por la Comisión tanto en el Libro Blanco como en las *Directrices éticas para una IA fiable* de su Grupo de Expertos en IA⁷⁰¹.

⁷⁰⁰ Para HERRERA DE LAS HERAS, R., «El nuevo marco jurídico de la inteligencia artificial: las novedades del Reglamento Europeo», *Actualidad Civil*, n° 5, Sección Derecho digital, Mayo 2024, LA LEY 18067/2024, p. 15/27, el hecho de que el RIA sea de aplicación directa y conlleve una armonización en diversos aspectos y un marco común para todos los Estados miembros, unido al denominado efecto Bruselas, hacen que se trate de una normativa pionera en el mundo que condicionará la futura normativa en materia de IA de otras regiones y países del mundo.

⁷⁰¹ Un recorrido cronológico de todas las iniciativas sobre IA que se han presentado en el marco de la UE en los últimos años puede verse en COCA PAYERAS, M., «Las iniciativas de la Unión Europea sobre inteligencia artificial: De la persona electrónica al difícil equilibrio entre la necesidad de impulsarla y evitar sus riesgos», *Revista de Derecho Civil*, vol. 10, n. 2, 2023, pp. 3-40. También en ÁLVAREZ CUESTA, H. *op.cit.*, p. 385 y ss. Hay que precisar que de las tres *Resoluciones del Parlamento Europeo, de 20 de octubre de 2020* mencionadas, la que mayor preocupación demuestra por evitar la discriminación algorítmica es la *Resolución sobre aspectos éticos*, donde se configura la ausencia de sesgos y discriminación como un principio ético de obligado cumplimiento para las tecnologías de alto riesgo y se concibe la discriminación como una de las posibles causas en la producción de daños por tales sistemas. Vid., concretamente, art. 4, apartado n). Señala LAZCOZ MORATINOS, G., «Análisis de la propuesta de Reglamento sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas», *IUS ET SCIENTIA: Revista electrónica de*

En estos documentos se contempla la discriminación algorítmica como un riesgo muy relevante frente al que ha de establecerse un marco normativo de *tutela preventiva*. Por tanto, se propone la supervisión y prevención de daños *ex ante*, comprobando que los sistemas de IA no reproduzcan discriminaciones sociales ya existentes, ni creen nuevos tipos o formas de discriminación.

2.1.1. El riesgo de discriminación de los sistemas en el RIA

La propuesta original del RIA lanzada por la Comisión en abril de 2021, constituyó el primer intento mundial de regulación horizontal de los sistemas de IA cuya finalidad era introducir un marco jurídico dirigido a conseguir una IA segura, fiable y respetuosa con los derechos fundamentales y los principios éticos.

En aquel texto inicial se prestaba una atención especial a la discriminación algorítmica, a la que se hacía referencia expresa. En su Exposición de Motivos, la Comisión declaraba que la Propuesta venía a «completar el Derecho de la Unión vigente en materia de no discriminación al establecer requisitos específicos que tienen por objeto reducir al mínimo el riesgo de discriminación algorítmica, en particular en lo tocante al diseño y la calidad de los conjuntos de datos empleados para desarrollar sistemas de IA», que iban acompañados de obligaciones sobre la realización de pruebas, la gestión de riesgos, la documentación y la vigilancia humana durante todo el ciclo de vida de los sistemas, con las que se pretendía garantizar un *control previo* o *ex ante* de los sistemas para evitar la vulneración de derechos y reducir al mínimo el riesgo de adoptar decisiones asistidas por IA erróneas o sesgadas en esferas críticas como la educación y la formación, el empleo, el acceso a determinados servicios y prestaciones esenciales públicas y privadas, la aplicación de la ley y el poder judicial⁷⁰².

A pesar de que el texto original ha sufrido importantes modificaciones a lo largo de su tramitación⁷⁰³, sigue manteniendo el enfoque basado en el riesgo

Derecho y Ciencia, vol. 6, n. 2, 2020, p. 36, que el derecho a la no discriminación está presente en toda la propuesta como uno de los pilares básicos del Derecho de la Unión.

⁷⁰² Vid., ap. 1.2 y 3.5. Como aprecia HUERGO LORA, A., «El proyecto de Reglamento sobre la Inteligencia Artificial», *cit.*, tales requisitos son objetivos máximos que pueden conseguirse de muchas maneras y con niveles de intensidad diferentes. Por ello, en cada caso habrá que buscar el mejor modo de cumplirlos. Además, no hay una única forma de hacerlo ya que se puede aspirar a un nivel máximo, medio o mínimo de calidad o de seguridad.

⁷⁰³ Las modificaciones más importantes del texto inicial se realizaron principalmente en virtud de la adopción de la Orientación General del Consejo aprobada el 6 de diciembre de 2022 por el Consejo de Transporte, Telecomunicaciones y Energía y se dirigieron a incorporar los sistemas de IA de propósito general (no contemplados en la propuesta originaria) y a modificar el alcance

que pueden implicar los distintos sistemas de IA (o el *uso* que se haga de ellos) de menoscabar la salud, seguridad o derechos fundamentales de las personas, diferenciando niveles de riesgo y clasificándolos en riesgo inaceptable, alto riesgo y riesgo bajo o mínimo⁷⁰⁴. Según la intensidad del riesgo y el bien jurídico a proteger se proponen una serie de controles *ex ante* y *ex post*. A mayor riesgo, mayores controles o requisitos para su puesta en circulación e incluso prohibición. A menor riesgo, más facilidades para su puesta en servicio.

Desde un primer momento se considera necesario, en primer lugar, prohibir determinadas prácticas de IA que conlleven un riesgo inaceptable (art.5); en segundo lugar, definir los requisitos que deben cumplir los sistemas de IA de alto riesgo (art. 8 a 15) que se traducen en estrictas obligaciones para proveedores, responsables del despliegue y otras partes (art. 16 a 27); en tercer lugar, fomentar para los sistemas de bajo riesgo la adopción de códigos de conducta dirigidos a que cumplan voluntariamente los requisitos obligatorios para los de alto riesgo (art. 95); y por último, imponer obligaciones específicas de transparencia a determinados sistemas de IA al margen de si resultan de alto riesgo (art. 50)⁷⁰⁵.

Por lo que se refiere al peligro de que los sistemas de IA contengan sesgos y generen discriminación injusta, desde la redacción original de la propuesta ha sido tenido muy en cuenta para clasificar un determinado tipo de sistema dentro de uno u otro nivel de riesgo. Así, por ejemplo, para prohibir los sistemas de *social scoring* o puntuación y clasificación ciudadana bajo ciertas condiciones⁷⁰⁶, los de categorización biométrica que clasifiquen individualmente

de la propia definición de IA. Paralelamente, el texto siguió su tramitación en el Parlamento donde se debatieron numerosas enmiendas y tras ello se sometió a votación en la Comisión de Mercado Interior y Protección del Consumidor y la Comisión de Libertades Civiles el 11 de mayo de 2023. El 9 de diciembre de 2023, el Consejo y el Parlamento alcanzaron un acuerdo provisional sobre el RIA. El 13 de marzo de 2024 el Parlamento aprobó en primera lectura el texto y el 21 de mayo el Consejo también le dio luz verde. Su versión definitiva se publicó en el DOUE de 12 de julio de 2024, con entrada en vigor el 1 de agosto y aplicación desde el 2 de agosto de 2026, si bien algunas previsiones son aplicables en plazos distintos.

⁷⁰⁴ Para HERRERA DE LAS HERAS, R., «El nuevo marco jurídico de la inteligencia artificial ...», *cit.*, p. 14/27, se trata de un «sistema loable y que será dinámico, puesto que la clasificación de los sistemas podrá ser modificada durante su aplicación». En sentido similar, PEGUERA POCH, M., «Perspectiva y categorización del riesgo ...», *cit.*, p. 926, mantiene que esta clasificación, no es ajena a las influencias de la industria y el activismo civil y está llamada a sufrir modificaciones.

⁷⁰⁵ Vid., el apartado 2.5.4.1 del Capítulo III. En particular, es preciso notificar a las personas físicas que están interactuando con un sistema de IA, salvo que sea evidente por las circunstancias y el contexto de uso, e informarlas cuando estén expuestas a un sistema de reconocimiento de emociones o a un sistema de categorización biométrica. Esta información y estas notificaciones deben facilitarse en formatos accesibles para las personas con discapacidad.

⁷⁰⁶ Se advierte en el Considerando 31 que tales sistemas pueden tener resultados discriminatorios y abocar a la exclusión a determinados grupos, menoscabando el derecho a la dignidad y a la no

a las personas físicas sobre la base de sus datos biométricos para deducir o inferir su raza, opiniones políticas, afiliación sindical, convicciones religiosas o filosóficas, vida sexual u orientación sexual, y los de identificación biométrica remota «en tiempo real» en espacios de acceso público con fines de aplicación de la ley (salvo las excepciones previstas), respecto de los que se advierte expresamente que, debido a sus imprecisiones técnicas, pueden provocar resultados sesgados y tener consecuencias discriminatorias, lo que cobra especial importancia en relación a la edad, la etnia, la raza, el sexo o la discapacidad⁷⁰⁷.

Se prohíben también los sistemas de evaluación de riesgos que se utilicen para evaluar o predecir el riesgo de que una persona cometa un delito, basándose únicamente en la elaboración de perfiles o en la evaluación de sus rasgos y características de personalidad. Aunque no se diga expresamente, el riesgo de discriminar en estos casos también es elevado y junto con la necesidad de respetar la presunción de inocencia, justifica esta nueva prohibición que no se contemplaba en el texto inicial. No obstante, se excluyen de esta prohibición «los sistemas de IA utilizados para *apoyar la evaluación humana* de la implicación de una persona en una actividad delictiva, que ya se basa en hechos objetivos y verificables directamente relacionados con una actividad delictiva» (art.5.1.d)⁷⁰⁸.

Igualmente, quedan prohibidos los sistemas de IA para inferir emociones en el lugar de trabajo o en centros educativos, debido a las dudas que generan

discriminación y los valores de igualdad y justicia. Se prohíben, en particular, los destinados a evaluar o clasificar a personas físicas o colectivos a lo largo de un cierto tiempo a partir de su comportamiento social o de «características personales o de su personalidad conocidas, inferidas o predichas», siempre que la puntuación resultante ocasione: bien un trato perjudicial o desfavorable a personas concretas o colectivos en contextos sociales que no guarden relación con los contextos donde se generaron o recabaron los datos originalmente; o bien un trato perjudicial o desfavorable para las personas afectadas, que sea injustificado o desproporcionado con respecto a su comportamiento social o la gravedad de este. Criticando la falta de concreción del precepto para delimitar la práctica prohibida y augurando que el mismo generará muchas dudas interpretativas y con ellas inseguridad jurídica, PEGUERA POCH, M., «Perspectiva y categorización del riesgo ...», *cit.*, p. 933. Para este autor, «del tenor del precepto cabría colegir que si el trato perjudicial, o incluso meramente desfavorable, se produce en el mismo contexto en el que se recabaron los datos –por ejemplo el lugar de trabajo– el sistema no quedaría afectado por la prohibición, salvo que el trato resulte a su vez injustificado o desproporcionado. Por otra parte, no se proporcionan elementos para valorar el carácter injustificado o desproporcionado, a la vez que se da a entender que ese perjuicio podría resultar aceptable si guarda proporción con el comportamiento social observado, cuestión que suscita algunos interrogantes».

⁷⁰⁷ Vid., Considerando 32.

⁷⁰⁸ Dispone el Considerando 42 que, en consonancia con la presunción de inocencia, las personas deben ser juzgadas basándose en su comportamiento real y nunca a partir de comportamientos predichos por una IA basados únicamente en la elaboración de sus perfiles, en los rasgos o características de su personalidad, como la nacionalidad, el lugar de nacimiento, el lugar de residencia, el número de hijos, el nivel de endeudamiento o el tipo de vehículo, *sin una evaluación humana y sin que exista una sospecha razonable, basada en hechos objetivos comprobables, de que dicha persona está implicada en una actividad delictiva*.

la fiabilidad y la base científica para la detección o deducción de emociones que se manifiestan de modo diverso en distintas culturas y colectivos y que pueden conducir a resultados discriminatorios⁷⁰⁹. Se excluyen del ámbito de la prohibición aquellos sistemas que tengan fines exclusivamente médicos o de seguridad.

Por lo que se refiere en concreto al ámbito laboral, se reconoce a lo largo del Reglamento que los sistemas de IA utilizados sobre todo para la contratación y selección de personal, pueden perpetuar patrones históricos de discriminación, por ejemplo contra las mujeres, ciertos grupos de edad, personas con discapacidad o personas de orígenes raciales o étnicos concretos o con una orientación sexual determinada, y en base a ello se califican como de alto riesgo⁷¹⁰. Y lo mismo ocurre con los que se usan para evaluar la calificación crediticia o solvencia de personas físicas (que deciden si las personas pueden acceder a recursos financieros o servicios esenciales como la vivienda, la electricidad y los servicios de telecomunicaciones), los cuales pueden discriminar a personas o grupos y perpetuar patrones históricos de discriminación o generar nuevas formas de discriminar⁷¹¹. Su potencialidad de afectar a éste y otros derechos fundamentales conlleva su calificación como sistemas de alto riesgo.

⁷⁰⁹ Así se justifica su prohibición en el Considerando 44. Se explica en el Considerando 18 que el «sistema de reconocimiento de emociones» debe definirse como un sistema de IA destinado a distinguir o deducir las emociones o las intenciones de las personas físicas a partir de sus datos biométricos y se refiere a emociones o intenciones como la felicidad, la tristeza, la indignación, la sorpresa, el asco, el apuro, el entusiasmo, la vergüenza, el desprecio, la satisfacción y la diversión, sin incluir los estados físicos, como el dolor o el cansancio (como, por ejemplo, los sistemas utilizados para detectar el cansancio de los pilotos o conductores profesionales con el fin de evitar accidentes). Tampoco incluye la mera detección de expresiones, gestos o movimientos que resulten obvios, salvo que se utilicen para distinguir o deducir emociones. Esas expresiones pueden ser expresiones faciales básicas, como un ceño fruncido o una sonrisa; gestos como el movimiento de las manos, los brazos o la cabeza, o características de la voz de una persona, como una voz alzada o un susurro.

⁷¹⁰ Considerando 57.

⁷¹¹ Considerando 58. También se mencionan y consideran de alto riesgo los sistemas de evaluación de riesgos y la fijación de precios en seguros de vida y de salud porque pueden afectar de un modo considerable a los medios de subsistencia de las personas y, si no se diseñan, desarrollan y utilizan debidamente, pueden vulnerar sus derechos fundamentales y pueden tener graves consecuencias para la vida y la salud de las personas, como la exclusión financiera y la discriminación. No obstante, se precisa en este Considerando (y en el Anexo III, ap.5) que no deben considerarse de alto riesgo los sistemas para detectar fraudes en la oferta de servicios financieros y para calcular los requisitos de capital de las entidades de crédito y las empresas de seguros. Por otro lado, desaparece en la versión final del RIA la exención contenida en el Considerando 37 de la propuesta inicial respecto a los sistemas de IA destinados a evaluar la solvencia y la calificación crediticia puestos en servicio por proveedores a pequeña escala para su propio uso.

2.1.2. Breve referencia a los modelos de IA de uso general

La propuesta inicial del RIA pretendía regular solo determinadas prácticas y casos específicos de uso de sistemas de IA y no establecía nada sobre los sistemas de *propósito o uso general* que podían ser utilizados para múltiples finalidades y en muy diversos sectores. No obstante, las herramientas de IA generativa (y más concretamente la irrupción en el mercado de ChatGpt en noviembre de 2022) obligaron a incorporar previsiones al respecto. Así, en su Orientación General de diciembre de 2022, el Consejo consideró necesario, por un lado, aclarar que los sistemas de IA de uso general eran sistemas concebidos por el proveedor para desempeñar funciones de aplicación general, como el reconocimiento de imágenes y de voz, en diversos contextos; y por otro lado, que podían utilizarse como sistemas de alto riesgo por sí solos o ser componentes de otros sistemas de alto riesgo. Además, teniendo en cuenta su naturaleza particular y a fin de garantizar un reparto equitativo de las responsabilidades a lo largo de la cadena de valor de la IA, estimó que dichos sistemas debían estar sujetos a requisitos y obligaciones proporcionados y más específicos⁷¹².

En la posición negociadora del Parlamento Europeo de 14 de junio de 2023, se incluyó una nueva subcategoría, la del *modelo fundacional*, término que si bien se emplea como sinónimo de los sistemas de propósito general, hace referencia a modelos previos y superiores a estos⁷¹³. Se definía como «*modelo de sistema de IA entrenado con un gran volumen de datos, diseñado para producir información de salida de carácter general y capaz de adaptarse a una amplia variedad de tareas diferentes*»⁷¹⁴. Además, se establecían una serie de obligaciones para los proveedores de dichos modelos, principalmente de transparencia, como revelar que el contenido había sido generado por una IA, diseñar el modelo para evitar que generase contenido ilegal o publicar resúmenes de los datos protegidos por derechos de autor utilizados para el entrenamiento⁷¹⁵.

⁷¹² Considerando 12 *quarter*.

⁷¹³ PEGUERA POCH, M., «La propuesta de Reglamento de la IA: una intervención legislativa insoslayable en contexto de incertidumbre», en M. Peguera Poch (coord.) *Perspectivas regulatorias de la inteligencia artificial en la Unión Europea*, Madrid, 2023, p. 160.

⁷¹⁴ Enmienda 168.

⁷¹⁵ Vid., Enmienda 399. El Considerando 107 de la versión definitiva del RIA explica lo siguiente en relación a esta obligación: «Con el fin de aumentar la transparencia en relación con los datos utilizados en el entrenamiento previo y el entrenamiento de los modelos de IA de uso general, incluidos los textos y los datos protegidos por el Derecho en materia de derechos de autor, procede que los proveedores de dichos modelos elaboren y pongan a disposición del público un resumen suficientemente detallado de los contenidos utilizados para el entrenamiento del modelo de IA de uso general. Este resumen debe tener debidamente en cuenta la necesidad de proteger los secretos comerciales y la información empresarial confidencial y, al mismo tiempo, debe ser exhaustivo en general en su alcance en vez de técnicamente detallado, a fin de facilitar que las partes con intereses

Concretamente, disponía el Parlamento que dichos modelos debían evaluar y mitigar los posibles riesgos y perjuicios mediante un diseño, unas pruebas y un análisis adecuados, aplicar medidas de gobernanza de datos –en particular, una evaluación de los sesgos– y cumplir requisitos de diseño técnico para garantizar niveles adecuados de rendimiento, previsibilidad, interpretabilidad, corregibilidad, seguridad y ciberseguridad, así como cumplir las normas medioambientales, estar sujetos a obligaciones en materia de información y elaborar toda la documentación técnica necesaria para que los posibles proveedores de etapas posteriores pudieran cumplir las obligaciones del Reglamento. Por tanto, se consideraba esencial la colaboración entre proveedores (del modelo de IA de uso general y del sistema de IA basado en él) para que estos últimos pudieran mitigar riesgos y que el propio modelo adoptara medidas para evitar que los errores o posibles sesgos discriminatorios del mismo se propagaran a los sistemas ulteriores de la cadena⁷¹⁶. Se aclaraba, no obstante, que el hecho de estar sometidos a estos requisitos y obligaciones específicas no implicaba que debieran considerarse como sistemas de alto riesgo⁷¹⁷.

Finalmente, la versión definitiva del RIA incluyó estos modelos de IA de propósito o uso general en su Capítulo V, distinguiendo entre aquellos que *presentan riesgos sistémicos*⁷¹⁸ (los cuales deben cumplir requisitos adicionales y más estrictos como realizar evaluaciones de los modelos, analizar y mitigar los riesgos sistémicos e informar sobre los incidentes) y aquellos que *no los presen-*

legítimos, incluidos los titulares de derechos de autor, ejerzan y hagan cumplir sus derechos en virtud del Derecho de la Unión, por ejemplo, enumerando los principales conjuntos o recopilaciones de datos que hayan servido para entrenar al modelo, como los grandes archivos de datos o bases de datos privados o públicos, y proporcionando una explicación descriptiva sobre otras fuentes de datos utilizadas. Conviene que la Oficina de IA proporcione un modelo para el resumen, que debe ser sencillo y eficaz y permitir que el proveedor proporcione el resumen requerido en forma descriptiva»

⁷¹⁶ PEGUERA POCH, M., «La propuesta de Reglamento de la IA...», *cit.*, p. 161.

⁷¹⁷ Enmienda 101. Se explicaba también en la Enmienda 100 que estos modelos podían ofrecerse también como servicio (*AI-as-a-Service*, *AIaaS*), por ejemplo a través del acceso a una API, en cuyo caso era necesaria la cooperación con los proveedores de etapas posteriores durante todo el tiempo de prestación del servicio para mitigar riesgos, salvo que el proveedor del modelo fundacional transfiriera el modelo de entrenamiento, así como información amplia y adecuada sobre los conjuntos de datos y el proceso de desarrollo del sistema, o restringiera el servicio, como el acceso a la API, de modo que el proveedor situado en una etapa posterior pudiera cumplir plenamente el Reglamento sin asistencia adicional del proveedor original del modelo fundacional.

⁷¹⁸ Los riesgos sistémicos aumentan con las capacidades y el alcance de los modelos, pueden surgir durante todo el ciclo de vida del modelo y se ven influidos por las condiciones de uso indebido, la fiabilidad del modelo, la equidad y la seguridad del modelo, el grado de autonomía del modelo, su acceso a herramientas, modalidades novedosas o combinadas, las estrategias de divulgación y distribución, la posibilidad de eliminar las salvaguardias y otros factores. Las reglas para clasificar estos modelos como modelos de IA de uso general con riesgo sistémico y el procedimiento para hacerlo, se establecen en los arts. 51 y 52.

tan (que están sujetos a requisitos más limitados, por ejemplo, con respecto a la transparencia a lo largo de la cadena de valor, el respeto a los derechos de autor o la publicación de resúmenes sobre el contenido usado para entrenar los modelos)⁷¹⁹. Por riesgo *sistémico* se entiende el «riesgo específico de las capacidades de gran impacto de los modelos de IA de uso general, que tienen unas repercusiones considerables en el mercado de la Unión debido a su alcance o a los efectos negativos reales o razonablemente previsibles en la salud pública, la seguridad, la seguridad pública, los derechos fundamentales o la sociedad en su conjunto, que puede propagarse a gran escala a lo largo de toda la cadena de valor» (art.3.65). Entre dichos riesgos sistémicos se menciona la *difusión de contenidos discriminatorios* y se recalca, de nuevo, la necesidad de prestar atención a las maneras en que estos modelos pueden dar lugar a *sesgos dañinos y discriminación* que entrañan riesgos para las personas, las comunidades o las sociedades⁷²⁰.

Asimismo, se define el «modelo de IA de uso general» como un «*modelo de IA, también uno entrenado con un gran volumen de datos utilizando la autosupervisión a gran escala, que presenta un grado considerable de generalidad y es capaz de realizar de manera competente una gran variedad de tareas distintas, independientemente de la manera en que el modelo se introduzca en el mercado, y que puede integrarse en diversos sistemas o aplicaciones posteriores, excepto los modelos de IA que se utilizan para actividades de investigación, desarrollo o creación de prototipos antes de su comercialización*» (art.3.63). Y el «sistema de IA de uso general» como aquel «*sistema de IA basado en un modelo de IA de uso general y que puede servir para diversos fines, tanto para su uso directo como para su integración en otros sistemas de IA*» (art. 3.66)⁷²¹.

⁷¹⁹ Estas obligaciones las recoge el art. 53. Además de las obligaciones enumeradas en dicho precepto, los proveedores de *modelos de IA de uso general con riesgo sistémico* deberán cumplir adicionalmente las del art. 55.

⁷²⁰ FERNÁNDEZ RIVAYA, J. y GARCÍA LUENGO, J. *Publicado el Reglamento europeo de Inteligencia Artificial: empieza la cuenta atrás para su completa entrada en vigor*, afirman que estos sistemas podrían entrañar riesgos sistémicos «si tienen capacidades de gran impacto o de impacto equivalente (se considera que los modelos de inteligencia artificial de uso general entrenados con una potencia informática total de más de 10²⁵ FLOPS –ChatGPT 4 o GEMINI– entrañan riesgos sistémicos). Dado que estos modelos potentes podrían causar accidentes graves o ser utilizados indebidamente para ciberataques de amplio alcance, el reglamento impone obligaciones adicionales de evaluación y mitigación del riesgo, de comunicación en caso de incidentes, y de protección de la ciberseguridad (artículo 55), pudiendo los proveedores de estos sistemas acudir también a códigos de buenas prácticas para demostrar el cumplimiento de estas obligaciones». Vid., https://www.garrigues.com/es_ES/garrigues-digital/publicado-reglamento-europeo-inteligencia-artificial-empieza-cuenta-atras (última visita el 28 de enero de 2025).

⁷²¹ En el Considerando 100 se añade que «Cuando un modelo de IA de uso general esté integrado en un sistema de IA o forme parte de él, este sistema debe considerarse un sistema de IA de uso general cuando, debido a esta integración, el sistema tenga la capacidad de servir a diversos fines. Un sistema de IA de uso general puede utilizarse directamente e integrarse en otros sistemas de IA».

Se explica en el Considerando 97 que debe establecerse una definición clara del concepto de «modelos de IA de uso general» y diferenciarse del concepto de «sistemas de IA» para garantizar la seguridad jurídica, y que dicha definición debe basarse en las características funcionales esenciales de un modelo de IA de uso general, en particular la generalidad y la capacidad de realizar una amplia variedad de tareas diferenciadas. Se aclara, por un lado, que los modelos de IA de uso general pueden introducirse en el mercado de diversas maneras, por ejemplo, a través de bibliotecas, interfaces de programación de aplicaciones (API), como descarga directa o como copia física y que pueden modificarse o perfeccionarse y transformarse en nuevos modelos; y por otro lado, que aunque los modelos de IA son componentes esenciales de los sistemas de IA, no constituyen por sí mismos sistemas de IA, ya que requieren que se les añadan otros componentes, como, por ejemplo, una interfaz de usuario, para convertirse en sistemas de IA. También se afirma que lo normal es que los modelos de IA suelen estar integrados en los sistemas de IA y formar parte de dichos sistemas. Ejemplo típico de modelos de IA de uso general son los *grandes modelos de IA generativa*, ya que permiten la generación flexible de contenidos, por ejemplo, en formato de texto, audio, imágenes o vídeo, que pueden adaptarse fácilmente a una amplia gama de tareas diferenciadas⁷²².

2.1.3. Los sistemas de alto riesgo en el RIA

2.1.3.1. Concepto de sistema de alto riesgo

Dejando a un lado los modelos de uso general, cuya evolución es incierta, nos detendremos brevemente en los sistemas de alto riesgo –clave de bóveda de la regulación europea– y en los múltiples requisitos que deben cumplir.

Se consideran de *alto riesgo* los sistemas que afecten negativamente a la seguridad o derechos fundamentales de las personas y se dividen en dos categorías (art.6)⁷²³:

- 1) Los que están destinados a ser utilizados como componente de seguridad de uno de los productos contemplados en la legislación de armonización de la UE enumerados en el Anexo I, o son en sí mismos uno de dichos productos (por ejemplo, máquinas; juguetes; embarcaciones de recreo y motos acuáticas; ascensores; equipos radioeléctricos; instalaciones de transporte por cable; productos sanitarios;

⁷²² Considerando 99.

⁷²³ Sobre los distintos sistemas en función del riesgo puede verse MARTÍN-CASALS, M., «Las propuestas de la Unión Europea para regular la responsabilidad civil ...», *cit.*, pp. 63-66.

aviación civil; sistema ferroviario; vehículos a motor; etc.), siempre que el producto del que el sistema de IA sea componente de seguridad o el propio sistema de IA como producto, deba someterse a una evaluación de la conformidad realizada por un organismo independiente para su introducción en el mercado o puesta en servicio con arreglo a los actos legislativos de armonización de la Unión enumerados en el anexo I.

- 2) Los que se recogen expresamente en el Anexo III del RIA (por ejemplo, los destinados a utilizarse en la identificación biométrica remota; los destinados a utilizarse como componente de seguridad en la gestión y el funcionamiento del tráfico rodado y el suministro de agua, gas, calefacción y electricidad; los destinados para determinar el acceso o la asignación de personas físicas a los centros de educación y formación profesional; los destinados para la contratación o selección, clasificación y filtrado de personas físicas para puestos vacantes; los destinados para evaluar la solvencia crediticia de personas físicas o establecer su clasificación crediticia; los destinados a ser utilizados por las autoridades con el fin de realizar evaluaciones de riesgo individuales de personas físicas para determinar el riesgo de que cometan infracciones penales o reincidan en su comisión; los destinados a ayudar a las autoridades a examinar solicitudes de asilo, visado y permisos de residencia; los destinados a ayudar a una autoridad judicial en la investigación e interpretación de los hechos y de la ley; etc.)⁷²⁴.

Los sistemas de IA independientes, es decir, aquellos que no son componentes de seguridad o productos en sí mismos, se clasifican como de alto riesgo si, a la luz de su *finalidad prevista*, presentan un *alto riesgo de ser perjudiciales para la salud y la seguridad o los derechos fundamentales de las personas*, teniendo en cuenta tanto la *gravedad* del posible perjuicio como la *probabilidad* de que se produzca, y se utilizan en los ámbitos especificados en el Reglamento⁷²⁵. Por tanto, son

⁷²⁴ Son básicamente los pertenecientes a ocho ámbitos específicos que deberán registrarse en una base de datos de la UE: identificación biométrica y categorización de personas físicas; gestión y explotación de infraestructuras críticas; educación y formación profesional; empleo, gestión de trabajadores y acceso al autoempleo, acceso y disfrute de servicios privados esenciales y servicios y prestaciones públicas; aplicación de la ley; gestión de la migración, el asilo y el control de fronteras; asistencia en la interpretación jurídica y aplicación de la ley.

⁷²⁵ Considerando 52. Se precisa, además, que «Para identificar dichos sistemas, se emplean la misma metodología y los mismos criterios previstos para la posible modificación futura de la lista de sistemas de IA de alto riesgo, que la Comisión debe estar facultada para adoptar, mediante actos delegados, a fin de tener en cuenta el rápido ritmo del desarrollo tecnológico, así como los posibles cambios en el uso de los sistemas de IA».

tres los criterios a tener en cuenta para considerarlos de alto riesgo: primero, la *finalidad* del sistema en sí misma; segundo, los *posibles daños* que puedan causar; y tercero, la *gravedad y probabilidad* de que estos puedan ocasionarse⁷²⁶.

No obstante, se aclara en el Considerando 53 que pueden existir casos en que los sistemas relativos a esos ámbitos predefinidos en el RIA no entrañen un riesgo considerable de causar perjuicios a intereses jurídicos protegidos, al no influir sustancialmente en la toma de decisiones. Se entenderá que no influyen sustancialmente en la toma de decisiones cuando no afecten al fondo, ni por consiguiente al resultado, de la toma de decisiones humana o automatizada.

Respecto a esta novedosa precisión que no se incluía en la propuesta originaria, con la que se pretende evitar una clasificación automática de los sistemas que podría no ser adecuada en algunos casos, aprecia con acierto COTINO HUESO que «la relevancia del resultado o decisión de un sistema de IA en la decisión final que se adopte por quienes lo usan, ha pasado a ser en el RIA un elemento esencial para determinar si el sistema es de alto riesgo y, con ello, [para] la aplicación misma del futuro Reglamento»; así, la cuestión de hasta qué punto la salida o resultado del sistema de IA es *determinante* para la decisión finalmente adoptada por el usuario, ya no es solo relevante a los efectos de aplicar las garantías del art. 22 RGPD, sino que ahora también es decisiva para determinar si un sistema de IA es considerado de alto riesgo y, en consecuencia, para la aplicación de todas las garantías previstas en el RIA.

Las condiciones para que pueda considerarse que un sistema no plantea un riesgo importante de causar un perjuicio a la salud, la seguridad o los derechos fundamentales de las personas, *al no influir sustancialmente en el resultado de la toma de decisiones*, las precisa el art. 6.3. Así será cuando se cumplan una o varias de las siguientes condiciones:

- a) *Que el sistema de IA tenga por objeto llevar a cabo una tarea de procedimiento limitada.* Podrían ser aquellos que transformen datos no estructurados en datos estructurados, que clasifiquen en categorías los documentos recibidos o detecten duplicados entre un gran número de aplicaciones. La naturaleza restringida y limitada de estas tareas hace que solo presenten riesgos limitados que no aumentan por su utilización en un contexto que el RIA especifique como de alto riesgo.
- b) *Que el sistema de IA tenga por objeto mejorar el resultado de una actividad humana previamente realizada, que pudiera ser pertinente a los efectos de dicho listado.* En estos casos, el sistema solo añadiría un nivel adicional

⁷²⁶ El art. 3.2 define el «riesgo» como «la combinación de la probabilidad de que se produzca un perjuicio y la gravedad de dicho perjuicio».

a la actividad humana entrañando un riesgo menor. Podrían ser los casos en que los sistemas se dirijan a mejorar el lenguaje utilizado en documentos ya redactados, por ejemplo, en lo referente al empleo de un tono profesional o de un registro lingüístico académico o a la adaptación del texto a una determinada comunicación de marca.

- c) *Que el sistema de IA tenga por objeto detectar patrones de toma de decisiones o desviaciones con respecto a patrones anteriores y no esté destinado a sustituir la evaluación humana previamente realizada sin una revisión humana adecuada, ni a influir en ella.* Aquí el riesgo sería menor por cuanto el sistema se utiliza tras la evaluación humana y no pretende sustituirla o influir en ella sin una revisión adecuada. Serían, por ejemplo, los sistemas que pueden utilizarse para comprobar a posteriori si un profesor puede haberse desviado de su patrón de calificación determinado, a fin de llamar la atención sobre posibles incoherencias o anomalías.
- d) *Que el sistema de IA tenga por objeto llevar a cabo una tarea preparatoria para una evaluación pertinente en los casos del anexo III.* Esta condición abarca, entre otras cosas, soluciones inteligentes para la gestión de archivos, lo que incluye funciones diversas como la indexación, la búsqueda, el tratamiento de texto y del habla o la vinculación de datos a otras fuentes de datos, o bien los sistemas para la traducción de documentos iniciales. En estos casos, la posible repercusión del resultado del sistema sería muy escasa y no representaría un riesgo para la evaluación posterior.

Sin perjuicio de lo anterior, se precisa que siempre se considerarán de alto riesgo los sistemas del anexo III cuando se dirijan a la *elaboración de perfiles de personas físicas en el sentido del art. 4.4 RGPD*. Se considera que estos casos presentan un riesgo significativo de menoscabar la salud, seguridad o derechos fundamentales de las personas.

Por último, se advierte en el Considerando 53 que la Comisión deberá, previa consulta al Comité Europeo de IA, proporcionar directrices que especifiquen la aplicación práctica de las condiciones en virtud de las cuales los sistemas del anexo III no se considerarán excepcionalmente como de alto riesgo, que habrán de completarse con una lista de ejemplos prácticos de casos de uso de sistemas de IA que sean o no de alto riesgo.

2.1.3.2. *Requisitos de los sistemas de alto riesgo*

Todos los sistemas considerados como de alto riesgo, además de implantar y mantener un *sistema de gestión de riesgos* para identificar las amenazas conocidas y previsibles y adoptar las medidas oportunas para evitar o reducir los

daños (arts. 8 y 9)⁷²⁷, deberán cumplir obligatoriamente una serie de requisitos referentes a la *calidad de los conjuntos de datos utilizados*, empleando datos de alta calidad, suficientemente pertinentes y representativos, libres de errores y completos en vista de la finalidad del sistema y tener las propiedades estadísticas adecuadas (art. 10), a la *documentación técnica actualizada* que deberá elaborarse antes de su introducción en el mercado o puesta en servicio y servirá para demostrar que el sistema cumple los requisitos del RIA⁷²⁸, a la *capacidad de registro automático de eventos*, debiendo implementar un registro de actividad que permita trazar los resultados del algoritmo y conservar toda la documentación del proceso para que pueda ser evaluada por las Autoridades (arts. 11 y 12)⁷²⁹, a la *transparencia y comunicación de información a los responsables del despliegue* de forma clara y adecuada (art. 13), a la *vigilancia humana*, debiendo diseñarse y desarrollarse de forma que una persona física pueda supervisar su funcionamiento y asegurarse de que se usan según lo previsto a lo largo de todo su ciclo de vida (art. 14)⁷³⁰, y a la *precisión, solidez y ciberseguridad*⁷³¹ (art. 15).

⁷²⁷ Se define en el art. 9.2 como «un proceso iterativo continuo planificado y ejecutado durante todo el ciclo de vida de un sistema de IA de alto riesgo, que requerirá revisiones y actualizaciones sistemáticas periódicas». Su objeto es detectar y mitigar los riesgos pertinentes de los sistemas de IA para la salud, la seguridad y los derechos fundamentales. Vid., sobre el sistema de gestión de riesgos el Considerando 65.

⁷²⁸ Su contenido mínimo se recoge en el Anexo IV.

⁷²⁹ En el Considerando 71 se explica que «Para permitir la trazabilidad de los sistemas de IA de alto riesgo, verificar si cumplen los requisitos previstos en el presente Reglamento, así como vigilar su funcionamiento y llevar a cabo la vigilancia poscomercialización, resulta esencial *disponer de información comprensible sobre el modo en que se han desarrollado y sobre su funcionamiento durante toda su vida útil*. A tal fin, es preciso *llevar registros y disponer de documentación técnica que contenga la información necesaria para evaluar si el sistema de IA en cuestión cumple los requisitos pertinentes y facilitar la vigilancia poscomercialización*. Dicha información debe incluir las características generales, las capacidades y las limitaciones del sistema y los algoritmos, datos y procesos de entrenamiento, prueba y validación empleados, así como documentación sobre el sistema de gestión de riesgos pertinente, elaborada de manera clara y completa. La documentación técnica debe mantenerse adecuadamente actualizada durante toda la vida útil del sistema de IA. Además, *los sistemas de IA de alto riesgo deben permitir técnicamente el registro automático de eventos, mediante archivos de registro, durante toda la vida útil del sistema*».

⁷³⁰ El proveedor deberá definir las medidas para ello antes de su introducción en el mercado o puesta en servicio y deberán *ser proporcionales a los riesgos, al nivel de autonomía y al contexto de uso del sistema de IA de alto riesgo*. Con ellas se debe garantizar, cuando proceda, que el sistema esté sujeto a limitaciones operativas que no se puedan desactivar, que responda al operador humano y que las personas físicas que lo supervisen posean las competencias, la formación y la autoridad necesarias para desempeñar esa función. También deberán incluir mecanismos para orientar e informar a dichas personas para que puedan tomar decisiones con conocimiento de causa acerca de si deben intervenir, cuándo hacerlo y de qué manera, o de detener el sistema si no funciona según lo previsto. Especial mención se hace en el art. 14.4.b) a que las personas que supervisan estos sistemas puedan ser conscientes del «sesgo de automatización», esto es, la tendencia de los humanos a confiar automáticamente o en exceso en la información de salida generada por un sistema de IA.

Además, como parte de las obligaciones a asumir por los proveedores, deberán asegurarse de que sus sistemas queden sometidos al *procedimiento de evaluación de la conformidad a que se refiere el art. 43 RIA antes de su introducción en el mercado o puesta en servicio* con el fin de demostrar que sus sistemas cumplen los requisitos anteriores y son altamente fiables (art. 16.f) ⁷³². Se entiende por «evaluación de la conformidad», el proceso por el que se *demuestra* si se han cumplido los requisitos establecidos en el capítulo II, sección 2, en relación con un sistema de IA de alto riesgo (art. 3.20).

Cuando dicha evaluación de la conformidad demuestre que los sistemas de IA cumplen los requisitos pertinentes, los proveedores elaborarán una *declaración UE de conformidad* con arreglo al art. 47 (art. 16.g) y colocarán el *marcado CE de conformidad* en el sistema de IA de alto riesgo o, cuando no sea posible, en su embalaje o en la documentación que lo acompañe, para indicar que respeta el Reglamento, de acuerdo con lo dispuesto en el artículo 48 (art. 16.h).

Cuando la evaluación de la conformidad exija la intervención de un organismo notificado, tras la misma, se expedirá un certificado (arts. 43 y 44).

Debe tenerse en cuenta que el RIA recoge dos sistemas para la evaluación de la conformidad: uno con intervención de un organismo notificado, y otro que implica un mero control interno. Como regla general, se aplicará el sistema interno y será el propio proveedor el que deberá llevar a cabo la evaluación de la conformidad de los sistemas bajo su propia responsabilidad (con la única excepción de los sistemas de IA que están destinados a utilizarse para la biometría). No obstante, como se explica en el Considerando 125, la complejidad de los sistemas de IA de alto riesgo y los riesgos asociados a ellos, implica que deba desarrollarse un procedimiento en el que participen orga-

Respecto a esta obligación de supervisión humana, la doctrina venía considerando que lo que pretendía el RIA era que los revisores humanos comprobaran las recomendaciones de los sistemas y no las aplicaran de forma rutinaria y acrítica, así como que tuvieran la capacidad y competencia para no adoptarlas una vez interpretadas y sopesadas. En este sentido, ÁLVAREZ CUESTA, H. *op.cit.*, p. 398. Para SORIANO ARNANZ, A., «La Propuesta de Reglamento de Inteligencia Artificial ...», *cit.*, p. 18, la intervención de un ser humano en las decisiones de los sistemas de IA (analizándolas y supervisándolas), teóricamente responde a la necesidad de no deshumanizar a las personas que son sometidas a procesos automatizados de tratamiento de datos y que no sean consideradas como objetos. Pero en la práctica, dicha supervisión humana aporta al proceso decisorio un grado de empatía que no tienen los sistemas de IA y garantiza que se tengan en cuenta las particularidades de cada sujeto y que, por tanto, se respete el debido proceso tecnológico.

⁷³¹ Sobre estos requisitos pueden verse también los Considerandos 74, 75 y 76.

⁷³² En opinión de algunos autores, esta evaluación del riesgo *ex ante* exigida por el art. 43 ofrece un buen equilibrio entre la facilidad de innovación y un alto nivel de protección proactiva de los derechos fundamentales. Así, ÁLVAREZ CUESTA, H. *op.cit.*, p. 400.

nismos notificados, denominado «*evaluación externa de la conformidad*», cuyo alcance se limitará, al menos en la primera fase de aplicación del RIA, a los sistemas de IA de alto riesgo que no están asociados a productos. Además, se prevé el nombramiento o constitución por cada Estado miembro de las denominadas «autoridades notificantes», que serán responsables de establecer y llevar a cabo los procedimientos necesarios para la evaluación, designación y notificación de los organismos de evaluación de la conformidad, así como de su supervisión (art. 28). Tanto los organismos notificados como las autoridades notificantes se regulan en la sección 4 del capítulo III. No obstante, como apunta la doctrina, toda esta estructura institucional parece que tendrá poca relevancia teniendo en cuenta que pocas veces será obligatoria su intervención, al menos por lo que se refiere a la mayor parte de los sistemas del anexo III⁷³³.

Por último, se establece la conveniencia de que un sistema de IA se someta a una nueva evaluación de la conformidad cada vez que se produzca un cambio («modificación sustancial»⁷³⁴) que pueda afectar al cumplimiento por su parte del presente Reglamento (por ejemplo, cambio del sistema operativo o de la arquitectura de software), o cambie la finalidad prevista del sistema, ya que en estos casos se considera que se trata de un nuevo sistema de IA (art. 43.4). Sin embargo, en el caso de los sistemas de IA que siguen «aprendiendo» después de su introducción en el mercado o puesta en servicio (es decir, aquellos que adaptan automáticamente el modo en que desempeñan sus funciones), es necesario establecer normas que indiquen que no deben considerarse modificaciones sustanciales los cambios en el algoritmo y en su funcionamiento que hayan sido predeterminados por el proveedor y se hayan evaluado en el momento de la evaluación de la conformidad.

Hecho esto, deberán *registrar* dichos sistemas de IA de alto riesgo *en una base de datos de la UE* que la Comisión gestionará con el propósito de redoblar la transparencia y vigilancia públicas y de fortalecer la supervisión *ex post* por parte de las autoridades competentes (art. 49 en relación con el art. 71). También los responsables del despliegue de sistemas de IA de alto riesgo que

⁷³³ En este sentido, PEGUERA POCH, M., «La propuesta de Reglamento de la IA...», *cit.*, p. 173. Y para SORIANO ARNANZ, A., «La Propuesta de Reglamento de Inteligencia Artificial...», *cit.*, pp. 20-21, la ausencia de intervención de un tercero en el mecanismo *ex ante* de control genera serias dudas acerca de la verdadera efectividad de la propuesta normativa.

⁷³⁴ Qué debe entenderse por «modificación sustancial» se establece en el art. 3.23: «un cambio en un sistema de IA tras su introducción en el mercado o puesta en servicio que no haya sido previsto o proyectado en la evaluación de la conformidad inicial realizada por el proveedor y a consecuencia del cual se vea afectado el cumplimiento por parte del sistema de IA de los requisitos establecidos en el capítulo II, sección 2, o que dé lugar a una modificación de la finalidad prevista para la que se haya evaluado el sistema de IA en cuestión».

sean autoridades, órganos u organismos públicos, o personas que actúen en su nombre (con excepción de las autoridades encargadas de la aplicación de la ley, del control fronterizo, de la inmigración o del asilo, y de las autoridades que sean responsables del despliegue de sistemas de IA de alto riesgo en el ámbito de las infraestructuras críticas), antes de utilizar un sistema de IA de alto riesgo enumerado en el anexo III, deberán registrarse en dicha base de datos, seleccionarán el sistema que tengan previsto utilizar y registrarán su utilización. Además, si constatan que el sistema que quieren utilizar no ha sido registrado en dicha base de datos, no lo utilizarán e informarán de ello al proveedor o distribuidor.

Junto con estas medidas, ya previstas en la propuesta inicial de RIA, en el acuerdo provisional alcanzado entre Consejo y Parlamento en diciembre de 2023, se estimó necesario incrementar la protección de los derechos de los ciudadanos. Para ello, se propuso incorporar la obligación de que los implementadores de sistemas de alto riesgo llevasen a cabo una *evaluación de impacto* en los derechos fundamentales antes de ponerlo en marcha, importante laguna de la que adolecía el texto original. Al mismo tiempo, aseguraron el reconocimiento del derecho de todos los ciudadanos a presentar quejas sobre los sistemas de IA y a recibir explicaciones respecto de aquellas decisiones basadas en ellos que afectasen a sus derechos fundamentales.

Por lo que se refiere concretamente a las evaluaciones de impacto algorítmico (conocidas también como AIA, por sus siglas en inglés –*Algorithmic Impact Assessment*–), ha sido el art. 27 el encargado de acoger finalmente esta nueva obligación para los sistemas de alto riesgo. Su objetivo es identificar riesgos específicos para los derechos de las personas y concretar las medidas a adoptar en caso de que se materialicen estos riesgos⁷³⁵.

Dicha obligación, que habrá de llevarse a cabo *antes de desplegar los sistemas*⁷³⁶ y que una vez realizada *deberá notificarse a la correspondiente autoridad de vigilancia del mercado*, recae sobre determinados responsables del despliegue:

⁷³⁵ Se desarrolla esta obligación en el Considerando 96. Como mantienen GERARDS, J./XENIDIS, R., *op.cit.*, p. 148, las evaluaciones de impacto se integran entre las estrategias *ex ante* o *preventivas* de la discriminación y consisten en realizar una evaluación o análisis del programa con el fin de identificar, de manera preventiva, la probabilidad de que una determinada decisión repercuta negativamente en el derecho a la igualdad y no discriminación. Dependiendo de los resultados de dicha evaluación, podría decidirse no recurrir a un algoritmo, o tener en cuenta riesgos específicos al desarrollarlo y al tomar decisiones con respecto a los conjuntos de datos que se utilizarán para entrenar o alimentar el algoritmo, o sobre el uso de las salidas del algoritmo en los eventuales procesos de toma de decisiones.

⁷³⁶ De hecho, precisa el apartado segundo que se aplicará *al primer uso* del sistema y que *deberá actualizarse* cuando el responsable del despliegue considere que durante su uso alguno de los elementos del apartado 1 hubiera cambiado.

por un lado, los que sean organismos de Derecho público respecto de todos los sistemas de IA que sean de alto riesgo; por otro lado, sobre los operadores privados que presten servicios públicos respecto de los sistemas de IA referidos a esos servicios públicos; y por último, (independientemente del carácter público o privado de dichas entidades) y por razón del objetivo de los sistemas, sobre los responsables del despliegue de sistemas de alto riesgo a que se refiere el punto 5 letras b) y c) del anexo III, que son: primero, los sistemas de IA destinados a ser utilizados para evaluar la solvencia de personas físicas o establecer su puntuación crediticia, a excepción de los utilizados para detectar fraudes financieros; y segundo, los destinados a ser utilizados para la evaluación de riesgos y la fijación de precios en relación con las personas físicas en el caso de los seguros de vida y de salud.

A través de dicha evaluación se deberán determinar los siguientes extremos (art. 27.1): a) los procesos del responsable del despliegue en los que se utilizará el sistema de IA de alto riesgo en consonancia con su finalidad prevista; b) una descripción del plazo de tiempo y la frecuencia en que se pretende utilizar el sistema; c) las categorías concretas de personas físicas y grupos de personas que probablemente se vean afectados por el uso del sistema en ese contexto específico; d) los riesgos de perjuicio específicos que probablemente afecten a los derechos fundamentales de dichas personas o grupos⁷³⁷; e) una descripción de la aplicación de medidas de supervisión humana, de acuerdo con las instrucciones de uso; y f) las medidas que deben adoptarse en caso de que dichos riesgos se materialicen, entre las que se incluyen, por ejemplo, sistemas de gobernanza para ese contexto de uso específico o los procedimientos de tramitación de reclamaciones y las vías de recurso. No obstante, si a resultados de la evaluación de impacto relativa a la protección de datos que debe llevarse a cabo en virtud de la normativa ad hoc cuando se tratan datos personales ya se cumpliera cualquiera de las obligaciones establecidas en el precepto analizado, la evaluación de impacto en los derechos fundamentales complementará a la relativa a la protección de datos. Para facilitar el cumplimiento de dicha obligación a los responsables del despliegue y reducir la carga administrativa que implica, la Oficina de IA elaborará un modelo de cuestionario al efecto⁷³⁸.

⁷³⁷ Al llevar a cabo esta evaluación, el responsable del despliegue debe tener en cuenta la información pertinente para una evaluación adecuada del impacto, lo que incluye, por ejemplo, la información facilitada por el proveedor del sistema en las instrucciones de uso en virtud del art. 13.

⁷³⁸ Añade HERRERA DE LAS HERAS, R., «El nuevo marco jurídico de la inteligencia artificial ...», *cit.*, p. 11/27, basándose en lo que ya venía exigiendo la Agencia Europea de los Derechos Fundamentales, que esas evaluaciones (que deberán desarrollarse en un futuro próximo) tendrán que tener en cuenta la totalidad de los derechos fundamentales que pueden verse afectados por el uso del sistema que se pretende implementar y deberán abordar las siguientes cuestiones: 1) La des-

Por otro lado, las obligaciones de los proveedores no acaban cuando introducen en el mercado o ponen en servicio sus sistemas, sino que una vez hecho esto, quedan sometidos a *obligaciones de vigilancia posteriores a su comercialización* (art.72) con las que también se trata de prevenir daños, o en su caso de corregirlos, como establece el art. 3.25⁷³⁹.

El sistema de vigilancia poscomercialización recopilará, documentará y analizará los datos pertinentes que puedan facilitar los responsables del despliegue o que puedan recopilarse a través de otras fuentes sobre el funcionamiento de los sistemas de IA de alto riesgo durante toda su vida útil (art. 72.2)⁷⁴⁰. Dicho sistema va dirigido a que los proveedores puedan garantizar una mejora de los sistemas en función de su experiencia, a que puedan introducir medidas correctoras si fueran necesarias y es fundamental para abordar de forma rápida y eficiente los riesgos derivados de los sistemas IA que siguen aprendiendo con el tiempo.

En suma, con dicho sistema de vigilancia *ex post* se trata de garantizar que los errores y sesgos no se introducen con el tiempo, a medida que el sistema continúa aprendiendo mientras funciona. Si una vez autorizada la implementación del algoritmo de aprendizaje automático (y aun siendo este explicable), no se supervisa su funcionamiento, resultados e impacto, dando por hecho que llegará a resultados justos, se podrían pasar por alto posibles errores derivados de actualizaciones del sistema, de los nuevos datos con los que interactúa o de otros factores relativos al contexto o la realidad social que sean cambiantes. Por tanto, es recomendable monitorizar en el tiempo el comportamiento del sistema con respecto a factores cuya variación pueda afectar a su funcionamiento futuro (por ejemplo, cambios en el contexto social, económico, organizacional o medioambiental)⁷⁴¹.

cripción del sistema, sus objetivos y su usabilidad; 2) La evaluación de los riesgos que pueda generar el sistema para los derechos fundamentales que puedan verse afectados; 3) Un análisis del grado de automatización y la complejidad del sistema; y 4) Los potenciales daños que se podrían producir y como prevenirlos.

⁷³⁹ El art. 3.25 define el «sistema de vigilancia poscomercialización» como «todas las actividades realizadas por los proveedores de sistemas de IA destinadas a recoger y examinar la experiencia obtenida con el uso de sistemas de IA que introducen en el mercado o ponen en servicio, con objeto de detectar la posible necesidad de aplicar inmediatamente cualquier tipo de *medida correctora o preventiva* que resulte necesaria». PEGUERA POCH, M., «La propuesta de Reglamento de la IA...», *cit.*, p. 177, destaca el papel de las autoridades nacionales de vigilancia del mercado en el control posterior a la comercialización de los sistemas.

⁷⁴⁰ Sigue diciendo el precepto citado que esta vigilancia poscomercialización, por un lado, incluirá un análisis de la interacción con otros sistemas de IA (incluidos otros dispositivos y *software*) cuando sea procedente, y por otro lado, que no abarcará los datos operativos sensibles de los responsables del despliegue que sean autoridades encargadas de la aplicación de la ley.

⁷⁴¹ Sobre esta necesidad de monitorizar los sistemas a lo largo de todo su ciclo de vida para evitar sesgos, errores y consiguientes daños, ya se manifestaba la mayoría de la doctrina. Por todos,

Asimismo, se exige a los proveedores que cuenten con un sistema para *comunicar a las autoridades pertinentes cualquier incidente grave* asociado al uso de sus sistemas de IA, entendido como un incidente o defecto de funcionamiento que tenga como consecuencia, directa o indirecta, el fallecimiento o daños graves para la salud de una persona, una alteración grave e irreversible de la gestión o el funcionamiento de infraestructuras críticas, el incumplimiento de obligaciones derivadas del Derecho de la Unión destinadas a proteger los derechos fundamentales o daños graves a la propiedad o al medio ambiente (art. 3.49). Así, establece el art. 73 que los proveedores notificarán cualquier incidente grave a las *autoridades de vigilancia del mercado* de los Estados miembros en los que se haya producido dicho incidente, notificación que deberá efectuarse inmediatamente después de que el proveedor haya establecido un vínculo causal entre el sistema de IA y el incidente grave o la posibilidad razonable de que exista dicho vínculo y, en cualquier caso, a más tardar quince días después de que el proveedor, o en su caso el responsable del despliegue, tengan conocimiento del incidente grave (arts. 73.1 y 2)⁷⁴².

Después de informar de un incidente grave, el proveedor realizará sin demora las investigaciones necesarias en relación con el mismo y el sistema de IA afectado, lo que incluirá una evaluación de riesgos del incidente y las medidas correctoras⁷⁴³. Además, deberá cooperar con las autoridades competentes y, en su caso, con el organismo notificado afectado durante dichas investigaciones, y no emprenderá acción alguna que suponga la modificación del sistema de IA afectado de un modo que pueda repercutir en cualquier evaluación posterior de las causas del incidente sin haber informado antes de dicha acción a las autoridades competentes.

Para finalizar, debe advertirse de que este marco jurídico se aplicará a los agentes públicos y privados *de dentro y fuera de la UE, en la medida en que el sistema*

consideraba NAVAS NAVARRO que dicho seguimiento del sistema implicaría la realización de informes de impacto periódicos y el deber de informar al usuario en consecuencia para que, en caso de descubrir sesgos o errores, se pudieran adoptar las medidas necesarias para eliminarlos y, en su caso, reparar los daños causados por ellos.

⁷⁴² Tal como precisa el art. 73.2 *in fine*, «[e]l plazo para la notificación a que se refiere el párrafo primero tendrá en cuenta la magnitud del incidente grave». Y dispone el art. 73.9 que en el plazo de siete días desde que reciba la notificación, la autoridad de vigilancia del mercado adoptará medidas adecuadas tal como se establece en el artículo 19 del Reglamento (UE) 2019/1020, y seguirá los procedimientos de notificación previstos en dicho Reglamento.

⁷⁴³ En virtud del art. 26.5, si son los responsables del despliegue los que detectan un incidente grave, «informarán inmediatamente de dicho incidente, en primer lugar, al proveedor y, a continuación, al importador o distribuidor y a la autoridad de vigilancia del mercado pertinente. En el caso de que el responsable del despliegue no consiga contactar con el proveedor, el artículo 73 se aplicará *mutatis mutandis*. Esta obligación no abarcará los datos operativos sensibles de los responsables del despliegue de sistemas de IA que sean autoridades encargadas de la aplicación de la ley».

de IA o modelo de uso general se introduzca en el mercado de la Unión, la información de salida generada por el sistema de IA se utilice en la Unión o su uso afecte a personas establecidas en ella. Se aplicará en todos los sectores y no afectará al Derecho de la UE, en particular, al derecho vigente en materia de protección de datos, protección de consumidores, derechos fundamentales, empleo, protección de los trabajadores y seguridad de los productos, a los que el RIA complementa⁷⁴⁴. Afectará tanto a los *proveedores* (por ejemplo, un programador informático que diseña una herramienta para el cribado de currículums) como a los *responsables del despliegue* de sistemas de IA de alto riesgo (por ejemplo, una empresa que compre esa herramienta), *importadores*, *distribuidores*, *fabricantes de productos* que introduzcan en el mercado o pongan en servicio un sistema de IA junto con su producto y con su propio nombre o marca comercial y *representantes autorizados*, pero no se aplicará a los *usos privados no profesionales* según señala el art. 2.10⁷⁴⁵. Por tanto, cuando un responsable del despliegue persona física utilice el sistema de IA para una actividad puramente personal de carácter no profesional, no quedará sometido a los deberes del RIA. Nótese que en la versión aprobada en mayo de 2024 desapareció la excepción en virtud de la cual dichos sujetos quedaban sometidos a la obligación de transparencia impuesta en el anterior art. 52, ahora nuevo art. 50⁷⁴⁶.

2.1.3.3. Especial referencia a los requisitos relativos a los conjuntos de datos

La importancia que adquieren los datos en los sistemas de IA y el hecho de que se considere de forma prácticamente unánime por los expertos que son la causa más frecuente de discriminación algorítmica, hace que debamos detenernos un poco más en las exigencias que al respecto establece el RIA.

Para empezar, se advierte de que uno de los mayores peligros que implica el uso de *data sets* inadecuados es el riesgo de discriminación por los sesgos

⁷⁴⁴ Considerando 9. Sobre las exclusiones del ámbito de aplicación del RIA, puede verse HERRERA DE LAS HERAS, R., «El nuevo marco jurídico de la inteligencia artificial ...», *cit.*, pp. 5 y ss.

⁷⁴⁵ *Inteligencia artificial: preguntas y respuestas* (1 de agosto de 2024) en https://ec.europa.eu/commission/presscorner/detail/es/qanda_21_1683 (última visita el 29 de enero de 2025).

⁷⁴⁶ A pesar de dicha exclusión, los usuarios no profesionales sí que se incluyen bajo el ámbito de aplicación subjetivo de la *Propuesta de Directiva sobre responsabilidad extracontractual en la IA* y, por tanto, podrán ser demandados cuando causen daños mediante el empleo del sistema de IA. Ahora bien, como veremos después, en estos casos la presunción de causalidad sólo podrá aplicarse si se dan las circunstancias recogidas en el art. 4.6, esto es, cuando el usuario haya alterado el sistema de modo que afecte a su seguridad o no haya tenido en cuenta las condiciones de funcionamiento del sistema pudiendo hacerlo. En el Considerando 29 de dicha Propuesta se establece que el aligeramiento de la carga de probar la causalidad no siempre resulta apropiado cuando el demandado no sea un usuario profesional, sino una persona que utiliza el sistema de IA para sus actividades privadas.

intrínsecos que pueden contener. Se afirma también, que si el sistema de IA no está entrenado con datos de buena calidad, puede señalar a personas de manera discriminatoria, incorrecta o injusta⁷⁴⁷.

Cuando se emplean técnicas que implican el entrenamiento de modelos (como el *machine learning*) los datos han de ser de *alta calidad* para que el sistema funcione correctamente y no origine algún tipo de discriminación prohibida por el Derecho de la Unión⁷⁴⁸. Para conseguir que los conjuntos de datos de entrenamiento, validación y prueba alcancen ese nivel, es preciso instaurar *prácticas adecuadas de gestión y gobernanza de datos*. Además, cuando se trate de datos personales, dichas prácticas deben incluir la *transparencia sobre el fin original de la recopilación de datos*⁷⁴⁹.

Según establece el art. 10.2, las prácticas de gobernanza y gestión de datos deben centrarse, en particular, en la elección de un diseño adecuado, la recogida y origen de los datos, las operaciones de tratamiento oportunas para la preparación de los datos (como la anotación, el etiquetado, la depuración, la actualización, el enriquecimiento y la agregación), la formulación de los supuestos pertinentes, fundamentalmente en lo que respecta a la información que, ateniéndose a ellos, los datos miden y representan, la evaluación previa de la disponibilidad, la cantidad y la adecuación de los conjuntos de datos necesarios, el examen de posibles sesgos y las medidas para detectarlos, prevenirlos y minimizarlos, y la detección de posibles lagunas o deficiencias pertinentes en los datos que impidan el cumplimiento del RIA y la forma de subsanarlas (art. 10.2. letras a-h).

Como puede observarse, se sigue un enfoque basado en la calidad del dato que es coherente con los principios de la normativa de protección de datos. Así, las exigencias de que la recogida, actualización o depuración de datos sean adecuadas, tratan de garantizar la fiabilidad y exactitud de la información utilizada por los sistemas. También se hace hincapié en que debe prestarse una atención especial a la mitigación de sesgos en los conjuntos de datos para evitar discriminaciones prohibidas (especialmente cuando los datos de salida influyan en la información de entrada de futuras operaciones –*bucles de retroalimentación*–), ya que los sesgos pueden ser inherentes a los conjuntos de

⁷⁴⁷ Considerando 59.

⁷⁴⁸ Cuando los sistemas de alto riesgo no empleen técnicas que impliquen el entrenamiento de modelos de IA, los apartados 2 a 5 del art. 10 se aplicarán únicamente a los conjuntos de datos de prueba (art. 10.6)

⁷⁴⁹ En la versión del RIA aprobada en mayo de 2024 se añadió un nuevo Considerando, el número 69, donde se insiste en que, cuando se utilicen datos personales, debe garantizarse su protección y el derecho a la intimidad a lo largo de todo el ciclo de vida de los sistemas y deben aplicarse los principios sobre los que se asienta la normativa de protección de datos, en particular, el de minimización de datos y protección desde el diseño y por defecto.

datos subyacentes y condicionar los resultados de los sistemas perpetuando y amplificando la discriminación, sobre todo cuando se utilizan datos históricos.

No obstante, el hecho de que el precepto haga referencia simplemente a aspectos generales de gestión y gobernanza de datos, sin precisar metodologías concretas, puede suponer ciertos desafíos en su aplicación práctica. Por tanto, sería conveniente un desarrollo normativo adicional o el establecimiento de estándares técnicos que clarifiquen cómo implementar estos principios.

Además del sometimiento a dichas prácticas, los conjuntos de datos de entrenamiento, validación y prueba deben cumplir otros requisitos o criterios de calidad: en particular, han de ser *pertinentes, lo suficientemente representativos y tener las propiedades estadísticas adecuadas*, también en lo que respecta a las personas o los grupos de personas en las que en un principio se usará el sistema de IA de alto riesgo; y deben estar *lo más libres de errores posible y ser completos* en vista de la finalidad prevista del sistema de IA (art. 10.3), sin que este requisito afecte al uso de técnicas de protección de la intimidad en el contexto del desarrollo y la prueba de sistemas de IA. No obstante, por lo que se refiere a esta última obligación, en muchos casos se advierte una tensión inherente entre ambos objetivos, ya que las técnicas de anonimización o agregación de datos pueden afectar a la precisión y representatividad de los conjuntos de datos. Por tanto, sería deseable un marco más detallado para conciliar estas exigencias sin comprometer ni la calidad del dato ni los derechos de los individuos.

Además, estos conjuntos de datos deben tener en cuenta, en la medida necesaria para su finalidad prevista, los rasgos, características o elementos particulares del entorno geográfico, contextual, conductual o funcional específico en el que se pretende utilizar el sistema de IA (art. 10.4).

Un aspecto relevante del precepto es que, con el fin de proteger los derechos de terceros frente a la discriminación que podría provocar el sesgo de los sistemas de IA, se exceptúa la prohibición general del art. 9.1 RGPD de tratar categorías especiales de datos personales, que había sido criticada porque dificultaba la detección de supuestos de discriminación algorítmica. Ahora, según dispone el RIA, los proveedores deben ser capaces de tratar también categorías especiales de datos personales, como cuestión de *interés público esencial* en el sentido del art. 9.2.g) RGPD y del art.10.2.g) Reglamento (UE) 2018/1725, para garantizar que el sesgo de los sistemas de IA de alto riesgo se vigile, detecte y corrija. En consecuencia, a tenor del art. 10.5, dicho tratamiento se permitirá de forma excepcional en la medida en que sea estrictamente necesario para garantizar la vigilancia, detección y corrección de los sesgos asociados a los sistemas de IA de alto riesgo y con todas las garantías para los derechos y libertades fundamentales. Además, para poder realizar

dicho tratamiento han de cumplirse todas las condiciones establecidas tanto en el RIA como en la normativa de protección de datos⁷⁵⁰.

Por último, es evidente que el acceso a datos de alta calidad será clave para poder desarrollar y evaluar los sistemas. En este sentido, los espacios comunes europeos de datos establecidos por la Comisión y la facilitación del intercambio de datos entre empresas y con los Gobiernos en favor del interés público, serán esenciales para brindar un acceso fiable, responsable y no discriminatorio a datos de alta calidad con los que entrenar, validar y probar los sistemas de IA.

2.1.3.4. *Las obligaciones de los sujetos de la cadena de valor*

El deber de cumplir los requisitos que impone el RIA y las obligaciones que estos implican se distribuye entre los diversos sujetos de la cadena de valor de los sistemas de alto riesgo. Aunque, como se vio *supra*, la mayoría recaen principalmente sobre los proveedores (art. 16), también se establecen deberes para los representantes autorizados (art. 22), importadores (art. 23), distribuidores (art.24) y responsables del despliegue (art. 26).

Si bien la enumeración y desarrollo de las múltiples obligaciones que impone el RIA a todos estos agentes de la cadena de valor excede con creces los límites de este trabajo, nos gustaría realizar algunas consideraciones generales sobre las mismas:

La necesidad de clarificar la asignación de responsabilidades y funciones entre estos sujetos, conllevó que el texto transaccional al que se llegó en diciembre de 2023 incluyera cambios en la propuesta inicial del RIA. Se añadieron algunas disposiciones adicionales en los arts. 13 y 14 a fin de conseguir una cooperación más eficaz entre proveedores y responsables del despliegue, se aclaró la relación entre las obligaciones del RIA y las ya vigentes en virtud de otros actos legislativos –como la normativa de protección de datos o la legislación sectorial sobre servicios financieros–⁷⁵¹ y se definieron con mayor claridad los supuestos en que otros agentes de la cadena de valor debían asumir las responsabilidades del proveedor.

Son los Considerandos 83 y siguientes del texto definitivo los que inciden en estas cuestiones. Tras reiterar la necesidad de aclarar la función y obligaciones específicas de cada operador de la cadena, se añade que, cuando alguno

⁷⁵⁰ Estas condiciones se han adicionado en la versión final del RIA y se recogen en las letras a)-f) del art. 10.5.

⁷⁵¹ Para un mayor detalle sobre la complementariedad del RIA respecto a estas otras normativas, pueden verse los Considerandos 9, 10 y 11.

desempeñe más de una función al mismo tiempo (por ejemplo, porque actúe como distribuidor e importador), deberán cumplir de forma acumulativa todas las obligaciones pertinentes asociadas a dichas funciones.

Los supuestos en que otros agentes de la cadena de valor deberán asumir las obligaciones del proveedor se recogen en el art. 25, que lleva por título «*Responsabilidades a lo largo de la cadena de valor de la IA*». Dispone dicho precepto que, cuando se den las circunstancias en él previstas, cualquier distribuidor, importador, responsable del despliegue o tercero será considerado proveedor de un sistema de IA de alto riesgo y estará sujeto a las obligaciones del proveedor previstas en el art. 16. Tales circunstancias se darán cuando:

- a) ponga su nombre o marca comercial en un sistema de IA de alto riesgo previamente introducido en el mercado o puesto en servicio, sin perjuicio de los acuerdos contractuales que estipulen que las obligaciones se asignan de otro modo;
- b) modifique sustancialmente un sistema de IA de alto riesgo que ya haya sido introducido en el mercado o puesto en servicio de tal manera que siga siendo un sistema de IA de alto riesgo con arreglo al artículo 6;
- c) modifique la finalidad prevista de un sistema de IA, incluido un sistema de IA de uso general, que no haya sido considerado de alto riesgo y ya haya sido introducido en el mercado o puesto en servicio, de tal manera que el sistema de IA en cuestión se convierta en un sistema de IA de alto riesgo de conformidad con el artículo 6.

Se hace hincapié en la obligación del proveedor inicial de cooperar con los nuevos proveedores y de facilitarles la información necesaria, el acceso técnico u otra asistencia con el fin de que puedan cumplir las obligaciones del Reglamento, en particular respecto de la evaluación de la conformidad de los sistemas de alto riesgo. Se exceptúa al proveedor inicial de cumplir dichas obligaciones en los casos en que haya indicado claramente que su sistema de IA no debe ser transformado en un sistema de IA de alto riesgo y, por tanto, no está sujeto a la obligación de facilitar la documentación (art. 25.2).

También los fabricantes de productos que incorporen el sistema de IA son destinatarios de distintas obligaciones. Dispone el art. 25.3 que «en el caso de los sistemas de IA de alto riesgo que sean componentes de seguridad de productos contemplados en la legislación de armonización de la Unión enumerados en la sección A del anexo I, *el fabricante del producto será considerado proveedor del sistema de IA de alto riesgo y estará sujeto a las obligaciones previstas en el artículo 16* en alguna de las siguientes circunstancias: a) que el sistema de IA de alto riesgo se introduzca en el mercado junto con el producto bajo el

nombre o la marca comercial del fabricante del producto; b) que el sistema de IA de alto riesgo se ponga en servicio bajo el nombre o la marca comercial del fabricante del producto después de que el producto haya sido introducido en el mercado. El fabricante debe garantizar, en particular, *que el sistema de IA integrado en el producto final cumpla los requisitos del RIA*⁷⁵².

2.2. EL CONVENIO MARCO DEL CONSEJO DE EUROPA SOBRE INTELIGENCIA ARTIFICIAL, DERECHOS HUMANOS, DEMOCRACIA Y ESTADO DE DERECHO

El 17 de mayo de 2024 el Consejo de Europa adoptó el primer tratado internacional jurídicamente vinculante destinado a garantizar el respeto de los derechos humanos, del Estado de derecho y las normas jurídicas democráticas en el uso de los sistemas de IA. Dicho Convenio fue el fruto de dos años de trabajo del Comité sobre IA que reunió para redactarlo a los 46 Estados miembros del Consejo de Europa, la Unión Europea y 11 Estados no miembros (Argentina, Australia, Canadá, Costa Rica, Estados Unidos, Israel, Japón, México, Perú, la Santa Sede y Uruguay), así como a representantes del sector privado, la sociedad civil y el mundo académico, que participaron como observadores. Se abrió a la firma el 5 de septiembre de 2024, momento en que la Vicepresidenta de Valores y Transparencia de la Comisión Europea lo firmó en nombre de la UE. Fue firmado también por Andorra, Georgia, Islandia, Noruega, la República de Moldavia, San Marino, Reino Unido, así como Israel y Estados Unidos, y queda abierto a la adhesión de países de todo el mundo⁷⁵³.

Este acuerdo representa un hito muy significativo en la regulación mundial de la IA siendo su objetivo prioritario el establecimiento de un marco jurídico que abarca todo el ciclo de vida de los sistemas de IA y que se dirige a prevenir y minimizar los riesgos que estos pueden plantear para los derechos humanos, la democracia y el Estado de derecho, a la vez que promueve la innovación responsable (art.1.1). Cubre el uso de sistemas de IA en el sector público –incluidas las compañías que actúan en su nombre– y en el sector privado (art. 3.1). En lo que respecta a los actores del sector privado, tienen la opción de aplicar directamente las obligaciones de la Convención o implementar medidas alternativas apropiadas para cumplir con sus disposiciones.

En sintonía con el RIA, a través del cual la UE pretende aplicar el Convenio, y en plena coherencia con él y con otros actos legislativos de la

⁷⁵² Considerando 87.

⁷⁵³ Su entrada en vigor se prevé para el primer día del mes siguiente a la expiración de un período de tres meses desde la fecha en que cinco signatarios, incluidos al menos tres Estados miembros del Consejo de Europa, lo hayan ratificado.

Unión, adopta un enfoque basado en el riesgo para diseñar, desarrollar, usar y decomisar sistemas de IA, que permite adaptar las obligaciones y medidas que impone a la naturaleza y magnitud de los riesgos asociados a cada sistema de IA, garantizando una regulación proporcional y efectiva (art. 1.2). Además, acoge la definición de sistema de IA prevista en el RIA e incluye una serie de conceptos que recuerdan a los ya establecidos por las instituciones europeas en muchos de sus documentos anteriores, como el de una IA centrada en el ser humano, confiable y segura basada en principios como el de *dignidad humana y autonomía individual, transparencia y supervisión* (incluida la transparencia en los contenidos generados por IA y en las interacciones con los sistemas de IA), *rendición de cuentas y responsabilidad* por los efectos adversos de los sistemas, *fiabilidad y confianza* en sus resultados, e *innovación segura* a través del establecimiento de *sandboxes* regulatorios⁷⁵⁴.

Especialmente relevantes a los efectos del presente trabajo son los *principios de igualdad y no discriminación*, así como el de *privacidad y protección de datos personales* previstos en sus arts. 10 y 11. Respecto al primero, se establece la obligación de las partes de adoptar o mantener medidas que garanticen que los sistemas de IA respeten la igualdad (con especial mención a la igualdad de género) y la prohibición de la discriminación, a lo largo de todo su ciclo de vida. Además, con su firma los Estados se comprometen «a adoptar o mantener medidas destinadas a superar las desigualdades a fin de lograr resultados imparciales, justos y equitativos, en consonancia con sus obligaciones nacionales e internacionales aplicables en materia de derechos humanos» (art. 10.2).

Asimismo, las partes tendrán que garantizar la disponibilidad de recursos legales para las víctimas de violaciones de los derechos humanos relacionadas con el uso de sistemas de IA, lo que habrá de hacerse a través de la adopción de medidas dirigidas a documentar la información sobre los sistemas que puedan afectar de forma significativa a los derechos humanos, a facilitar esa información a los organismos autorizados a acceder a ella, y a ponerla a disposición o comunicarla a las personas afectadas cuando proceda. Se debe garantizar que tal información sea suficiente para que las personas afectadas sean capaces de impugnar las decisiones adoptadas mediante el uso del sistema o basadas en gran parte en tal uso y, cuando sea pertinente y proceda, el uso mismo del sistema (art. 14).

Los Estados parte también tendrán que asegurar que los afectados cuenten con garantías procesales suficientes, incluida la notificación a todas las personas que interactúen con sistemas de IA de que están efectivamente interactuando con una máquina (art. 15).

⁷⁵⁴ Todos estos principios se recogen en el Capítulo III.

En cuanto a las obligaciones de gestión de riesgos, se establece en el art. 16 que las partes deberán adoptar o mantener «medidas para la *identificación, evaluación, prevención y mitigación de los riesgos* que plantean los sistemas de inteligencia artificial, teniendo en cuenta los efectos reales y potenciales para los derechos humanos, la democracia y el Estado de Derecho» y que esas medidas se graduarán y diferenciarán en función de una serie de parámetros recogidos en el apartado segundo. También tendrán que evaluar la necesidad de una moratoria, una prohibición u otras medidas apropiadas en relación con el uso de sistemas de IA, cuando sus riesgos puedan ser incompatibles con el respeto de los derechos humanos, el funcionamiento de la democracia o el Estado de Derecho (art. 16.3).

Por último, para garantizar su aplicación efectiva se establece un mecanismo de seguimiento en forma de Conferencia de las Partes y se exige que, para vigilar su cumplimiento, cada parte establezca mecanismos de supervisión independientes. No obstante, se incluyen exenciones en la aplicación del Convenio establecidas en los apartados 2 a 4 del art. 3. Así, las Partes no estarán obligadas a aplicarlo a actividades relacionadas con la protección de sus intereses de seguridad nacional, actividades de investigación y desarrollo relativas a sistemas de IA que aún no se hayan puesto a disposición para su uso, a menos que las pruebas o actividades similares se lleven a cabo de manera que puedan interferir en los derechos humanos, la democracia y el Estado de Derecho, y a materias relativas a defensa nacional (art.3.2, 3.3 y 3.4 respectivamente).

De todo lo afirmado puede concluirse que el Convenio Marco analizado es compatible y complementario a otras normas de la UE en materia de IA y protección de datos. Con su firma, la Unión refuerza su compromiso de alcanzar una IA respetuosa con los derechos fundamentales, que promueva la democracia y se desarrolle dentro de un marco legal y ético sólido y coherente.

2.3. LAS MEDIDAS DE PREVENCIÓN DEL DAÑO EN EL RGPD

Sin perjuicio de las medidas y requisitos previstos ahora en el RIA principalmente para los sistemas alto riesgo, en nuestro ordenamiento ya contábamos con algunas herramientas con las que hacer frente a los daños derivados de las decisiones automatizadas.

Uno de los instrumentos fundamentales para ello ha sido hasta ahora el RGPD. Así lo ha venido afirmando un buen número de autores que han tratado el tema y que consideran que se trata de la «regulación más importante sobre el binomio derechos fundamentales e IA»⁷⁵⁵. También para la AEPD, la

⁷⁵⁵ Así SÁEZ LARA, C., *op.cit.*, p. 51. Destacando también la relevancia del RGPD para hacer frente a la discriminación algorítmica, SORIANO ARNANZ, A. «Decisiones automatizadas y discrimi-

aplicación del RGPD y de las garantías que incorpora permiten minimizar el efecto multiplicador que conllevan las tecnologías de IA de las carencias éticas de la sociedad y que se reflejan en los datos históricos.

Ello no obsta para que también se hayan puesto de manifiesto las carencias de dicha normativa para hacer frente concretamente a la discriminación algorítmica en tanto en cuanto, como es lógico, su objetivo principal no es proteger el derecho a la igualdad y no discriminación sino el derecho fundamental a la protección de datos personales. No obstante, al margen de dichas críticas que después se expondrán, es innegable que el RGPD ofrece posibilidades para hacer frente a las decisiones automatizadas discriminatorias⁷⁵⁶. Entre otras, por las siguientes razones:

Por un lado, porque los principios sobre los que se asienta (establecidos en el art. 5 del RGPD), no sólo obligan a los responsables y encargados del tratamiento a garantizar la transparencia hacia el interesado sobre cómo y por qué se recogen y utilizan sus datos personales y a garantizar su exactitud, sino que también implican *indirectamente* la obligación de considerar cómo la forma en que un sistema procesa los datos puede afectar a los derechos a la igualdad y a la no discriminación⁷⁵⁷. Por tanto, no solo se debe tener en cuenta cómo se recogen y usan los datos, sino también la necesidad de evaluar el impacto que su tratamiento puede tener en los derechos fundamentales, incluso de manera indirecta.

Por lo que se refiere a la transparencia, nos remitimos a lo ya expuesto en el apartado 2 del capítulo III. Respecto a la exactitud de los datos, tal como explica la AEPD, en el caso de tratamientos que incorporan soluciones IA, la existencia de sesgos en los modelos de inferencia está íntimamente ligada con la exactitud o calidad del dato⁷⁵⁸. Para garantizar que los datos asociados a los interesados sean exactos (tanto los directamente recogidos como los inferidos), el responsable del tratamiento debe utilizar «*procedimientos matemáticos o estadísticos adecuados para la elaboración de perfiles*», siendo obligatorio demostrar y documentar que dichos procedimientos son precisos y, por tanto, estables y predecibles. Además, la Agencia informa de diversos factores que pueden influir sobre la exactitud de los datos inferidos. Entre ellos, menciona

nación: aproximación...», cit., p. 4/31 o RIVAS p. 454, para quien «uno de los elementos fundamentales para articular la protección frente a la discriminación algorítmica, hasta alcanzar un mayor desarrollo normativo en la protección frente a las decisiones automatizadas, es precisamente, el análisis del tratamiento de datos».

⁷⁵⁶ SÁEZ LARA, C., *op.cit.*, p. 58.

⁷⁵⁷ SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación: aproximación...», cit., p.21/31.

⁷⁵⁸ El término «calidad del dato» aparece en el RGPD en el art. 47.2.d.

la propia implementación del sistema de IA, errores de diseño o programación del modelo, errores o sesgos contenidos en los conjuntos de datos de entrenamiento o validación (deliberadamente introducidos o no) o la propia evolución sesgada del modelo, y señala que, para garantizar la fidelidad e integridad del conjunto de datos de entrenamiento, se han de emplear métricas y técnicas de depuración y trazabilidad⁷⁵⁹.

Por otro lado, porque su art. 9 prohíbe el tratamiento de *categorías especiales de datos personales* (también conocidos como *datos sensibles*) que prácticamente vienen a coincidir con las denominadas *categorías sospechosas* o *protegidas* en la normativa antidiscriminatoria⁷⁶⁰. En concreto, datos que revelen el origen étnico o racial, las opiniones políticas, religiosas o filosóficas, los datos genéticos o biomédicos y aquellos otros referentes a la vida u orientación sexuales de la persona. Así pues, cualquier sesgo en los resultados del sistema causado al tener en cuenta este tipo de datos de manera voluntaria o por error, supondría el incumplimiento de este artículo cuando no concorra causa legítima para su tratamiento, además del incumplimiento de la prohibición legal de no discriminar.

Concretamente en el ámbito de las decisiones automatizadas y la elaboración de perfiles, el art. 22.4 también prohíbe que su adopción se base en dichas categorías de datos personales, *salvo que se aplique el artículo 9, apartado 2, letra a) o g), y se hayan tomado medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado*. Explica con más detalle el Considerando 71, que los responsables del tratamiento están obligados a introducir las medidas necesarias para evitar efectos discriminatorios basados en categorías especiales de datos. Así, dispone que dichos sujetos deberán «[...] aplicar medidas técnicas y organizativas apropiadas para garantizar, en particular, que se corrijan los factores que introducen inexactitudes en los datos personales y se reduce al máximo el riesgo de error; asegurar los datos personales de forma que se tengan en cuenta los posibles riesgos para los intereses y derechos del interesado y se impidan, entre otras cosas, efectos discriminatorios en las personas físicas por motivos de raza u origen étnico, opiniones políticas, religión o creencias, afiliación sindical, condición genética o estado de salud u orientación sexual, o que den lugar a medidas que produzcan tal efecto». De aquí se infiere que la noción de igualdad en el diseño se encuentra también indirectamente incluida en el marco normativo europeo en materia de protección de datos.

⁷⁵⁹ Informe sobre la Adecuación del RGPD a la IA, cit., p. 36/53.

⁷⁶⁰ Salvo en determinadas circunstancias mencionadas en el ap. 2, dispone el art. 9.1 que «Quedan prohibidos el tratamiento de datos personales que revelen el origen étnico o racial, las opiniones políticas, las convicciones religiosas o filosóficas, o la afiliación sindical, y el tratamiento de datos genéticos, datos biométricos dirigidos a identificar de manera unívoca a una persona física, datos relativos a la salud o datos relativos a la vida sexual o la orientación sexual de una persona física».

Además, como ya se explicó con detalle, el art. 22 RGPD consagra el derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado de datos personales, incluida la elaboración de perfiles, que produzca efectos jurídicos en la persona o le afecte significativamente de manera similar. Si bien se prevén ciertas excepciones a esta prohibición, en general, las decisiones automatizadas deben basarse en el consentimiento explícito del interesado y deben existir medidas adecuadas para salvaguardar sus derechos, libertades e intereses legítimos, incluyendo al menos el derecho a obtener intervención humana, expresar su punto de vista y a impugnar la decisión, así como a obtener una explicación tras la adopción de la decisión⁷⁶¹. A nuestro modo de ver, estas garantías son particularmente relevantes tanto para detectar como para prevenir la discriminación algorítmica y permiten a los interesados desafiar las decisiones de los sistemas.

Pues bien, lo dispuesto en los arts.9.1 y 22.4 RGPD permite afirmar que, en la medida en que el RGPD trata de evitar que la toma de decisiones se base en categorías de datos que también están protegidas por el Derecho antidiscriminatorio, como la raza o la religión, se erige como un instrumento de prevención de la discriminación a través de la anticlasificación. No obstante, el hecho de que en el art. 9 no se contemplen algunas categorías especiales de datos como los relativos al sexo, discapacidad, edad y estatus socioeconómico (que sin embargo constituyen datos que revelan la pertenencia a grupos desaventajados cuyos miembros son más susceptibles a recibir un trato discriminatorio), refleja que el verdadero objetivo del RGPD no es evitar la discriminación sino proteger la privacidad⁷⁶².

Además, debe tenerse en cuenta que la prohibición de tratar categorías especiales de datos tiene excepciones. Por ello, cuando concurren una serie de circunstancias recogidas en el art. 9.2, el tratamiento de estos datos estará permitido siempre y cuando se cumplan ciertas condiciones. En concreto, se autoriza el tratamiento de datos sensibles *cuando sea necesario por razones de un interés público esencial*, sobre la base del Derecho de la Unión o de los Estados miembros (letra g).

⁷⁶¹ Sobre el art. 22 RGPD nos remitimos a lo dicho en el apartado 2.5.2.1 del Capítulo III.

⁷⁶² En este sentido, SORIANO ARNANZ, A., «Discriminación algorítmica: garantías ...», *cit.*, p. 151. No obstante, a juicio de esta autora, podría llegar a interpretarse que categorías de datos como el sexo o la discapacidad se incluyen en otras categorías que sí recoge el artículo 9, por ejemplo, dentro de las categorías de datos relativas a la vida sexual (sexo) y salud (discapacidad). No obstante, sigue diciendo, quedaría sin resolver la cuestión del estatus socioeconómico de aquellas personas cuyos datos se procesan, aunque el RGPD no ignora el riesgo que el procesamiento de este tipo de datos genera para aquellos que se encuentran en situación económica precaria y los recoge en el Considerando 75, donde se reconoce el riesgo que se genera para los derechos y libertades de las personas la evaluación de aspectos referidos a su situación económica.

Dicho fundamento es el esgrimido en el RIA para justificar el tratamiento de estas categorías especiales de datos a través de sistemas IA. Evitar la discriminación que podría provocar el sesgo de los sistemas de IA se considera una cuestión de orden público esencial en virtud de la cual los proveedores deben ser capaces de tratar ese tipo de datos «*con carácter excepcional, en la medida en que sea estrictamente necesario para garantizar la detección y corrección de los sesgos asociados a los sistemas de IA de alto riesgo, con sujeción a las garantías adecuadas para los derechos y libertades fundamentales de las personas físicas*»⁷⁶³ y tras la aplicación de todas las condiciones aplicables establecidas en el presente Reglamento, además de las condiciones establecidas en los Reglamentos (UE) 2016/679 y (UE) 2018/1725 y la Directiva (UE) 2016/680.⁷⁶⁴ Es el art. 10.5 el que se encarga de establecer las condiciones para que se produzca dicho tratamiento.

Por último, el *principio de responsabilidad proactiva* que inspira la normativa de protección de datos obliga a los responsables del tratamiento a implementar medidas técnicas y organizativas apropiadas para garantizar y poder demostrar que el tratamiento es conforme al Reglamento (art. 24.1 RGPD). En base a ello, deberán asegurar que los datos personales se utilizan respetando los derechos y libertades fundamentales de las personas (entre ellos, la igualdad y no discriminación) y adoptar políticas internas y medidas como la seudonimización y el cifrado de datos personales⁷⁶⁵.

Según señala el art. 24.3 RGPD, la adhesión a códigos de conducta o mecanismos de certificación podrán ser utilizados como elementos para demos-

⁷⁶³ En la Orientación General del Consejo sobre el RIA de 25 de noviembre de 2022, el texto del art. 10.5 mencionaba como salvaguardias adecuadas para los derechos y libertades fundamentales, el establecimiento de «*limitaciones técnicas a la reutilización y la utilización de las medidas de seguridad y protección de la privacidad más recientes, tales como la seudonimización o el cifrado, cuando la anonimización pueda afectar significativamente al objetivo perseguido*».

⁷⁶⁴ Vid., Considerando 70.

⁷⁶⁵ En la *Resolución del Parlamento Europeo sobre Macrodatos*, de 14 de marzo 2017, se dice en el Considerando 1, que las técnicas de seudonimización y cifrado pueden atenuar los riesgos anejos a la analítica de macrodatos y desempeñar un papel importante en la garantía de la privacidad del interesado, fomentando asimismo la innovación y el crecimiento económico. Téngase en cuenta que las técnicas preventivas que permiten, entre otras cosas, la seudonimización y anonimización de datos, han sido avaladas por el GT29 en sus *Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679*, p. 36, como mecanismos para evitar la vulneración del derecho a la protección de datos. Para HERRERA DE LAS HERAS, R., «Protección de datos e inteligencia artificial», *cit.*, p. 658, la anonimización de datos en una herramienta ideal para el tratamiento de datos por aplicaciones que se basen en el uso de la IA. Cuestión distinta y responsabilidad de los encargados del tratamiento de datos será garantizar esa anonimización durante todo el proceso, de principio a fin. En ocasiones es una técnica compleja, pero es una de las principales herramientas para hacer compatibles la privacidad y la IA. Pueden verse las diferencias entre anonimización y seudonimización, en <https://www.aepd.es/prensa-y-comunicacion/blog/anonimizacion-y-seudonimizacion> (última consulta el 30 de enero de 2025).

trar el cumplimiento de las obligaciones por parte del responsable del tratamiento. También las evaluaciones de impacto se sitúan entre las medidas que los responsables deben aplicar para garantizar que sus tratamientos respetan el RGPD y estar en condiciones de demostrarlo. El art. 35 RGPD requiere que los responsables las realicen cuando del tratamiento de datos pueda resultar un riesgo elevado para los derechos y libertades de las personas físicas. En tales casos, se han de evaluar los riesgos que plantea el tratamiento de datos –especialmente en el caso de decisiones automatizadas y la elaboración de perfiles– e implementar medidas para mitigarlos.

En definitiva, puede afirmarse que a través de sus principios y disposiciones, el RGPD proporciona un marco robusto que puede ayudar a prevenir la discriminación algorítmica. Los principios de tratamiento justo y transparente de los datos, el derecho a no ser objeto de decisiones automatizadas, el principio de responsabilidad proactiva y las evaluaciones de impacto, son mecanismos que pueden contribuir a identificar y mitigar *ex ante* los riesgos de discriminación inherentes a los sistemas. Ahora bien, la implementación efectiva de estas medidas requiere un compromiso real por parte de las organizaciones y un adecuado control por parte de las autoridades competentes que, como se dirá en el siguiente apartado, está siendo limitado en la práctica.

2.3.1. Las carencias del RGPD para luchar contra la discriminación algorítmica

Aunque la normativa de protección de datos ofrece mecanismos útiles para detectar y prevenir, en cierta medida, situaciones de discriminación derivadas del tratamiento de datos personales, no es menos cierto que también presenta algunas deficiencias. Ello se debe, principalmente, a que el RGPD no es un instrumento dirigido a tutelar el derecho de igualdad de trato sino el derecho a la privacidad.

Algunas de las razones que la doctrina ha esgrimido para criticar la eficacia del RGPD como herramienta frente a las decisiones algorítmicas discriminatorias han sido las siguientes⁷⁶⁶:

⁷⁶⁶ Con más detalle las expone SORIANO ARNANZ en sus trabajos «Decisiones automatizadas y discriminación...», *cit.*, pp. 19 y ss. y «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, pp. 108 y ss. También destaca algunas de estas carencias, ALAMEDA CASTILLO, M.T., *op.cit.*, p. 32. Y también sobre las limitaciones del RGPD para hacer frente de forma suficiente a la discriminación, puede verse DIGITAL FUTURE SOCIETY, *La discriminación algorítmica en España: límites y potencial del marco legal*, 2022, pp. 12-13. Recuperado de https://digitalfuturesociety.com/app/uploads/2022/09/Discriminacion_algoritmica_Espana_marco_legal.pdf (último acceso el 1 de febrero de 2025).

En primer lugar, el hecho de que el sistema de protección de datos se articule sobre nociones de autonomía e individualidad complica la detección de errores sistémicos, que son los que se hallan en el origen de la discriminación algorítmica. Teniendo en cuenta que la discriminación algorítmica suele ser indirecta, que se imponga a los concretos interesados la carga de ejercitar los derechos que la norma confiere, impide detectar con carácter general si el algoritmo discrimina puesto que para conseguirlo será preciso casi siempre acceder a la totalidad del sistema y a los resultados que el mismo arroja para otras personas sometidas al procesamiento. Sólo si se tratara de un supuesto de discriminación directa (en que la pertenencia al grupo se considera de manera explícita) podría ser suficiente con que una persona interesada solicitara una explicación acerca de la lógica subyacente al sistema para que se pudiera detectar la discriminación. Además, el hecho de que tales derechos (información, acceso, rectificación, etc.) no estén diseñados para ejercer un control *ex ante*, sino que son mecanismos para que las personas puedan reaccionar *después* de que sus datos hayan sido procesados, merma la eficacia de esta normativa para la prevención de daños que puedan surgir de decisiones automatizadas.

En segundo lugar, el marco jurídico de protección de datos se centra en los datos introducidos en los sistemas (*inputs*) y no tanto en los resultados obtenidos (*outputs*). Es decir, la tutela se centra en la corrección y precisión de los datos introducidos inicialmente en el sistema y la forma en que estos son procesados, pero no se articulan mecanismos de protección frente a las inferencias obtenidas del procesamiento de datos. Así, en virtud de los arts. 12, 13 y 14 RGPD, el interesado gozaría del derecho a conocer los datos personales que ha recopilado la empresa usuaria del sistema de decisión automatizado, pero no tendría este derecho respecto a las inferencias realizadas mediante los datos disponibles, ni respecto a los perfiles que se elaboran sobre él⁷⁶⁷.

En tercer lugar, el hecho de que se establezcan prohibiciones de procesar ciertas categorías especiales de datos (arts. 9 y 22.4 RGPD), si bien minimiza el riesgo, no garantiza que el algoritmo no pueda inferir la pertenencia de la persona a una de ellas a partir de otros datos no sensibles e incluso no personales. De hecho, así ocurrió en algunos de los casos de discriminación algorítmica de mayor repercusión social en los últimos años, como el de Amazon. Por ello, la doctrina reconoce la insuficiencia en el ámbito del *machine learning* de las técnicas recogidas en el RGPD para anonimizar datos o despojarlos de rasgos personales, teniendo en cuenta que cualquier tipo de rasgo es susceptible de rastreo e inferencia para alcanzar un resultado de alta granularidad⁷⁶⁸.

⁷⁶⁷ GINÈS i FABRELLAS, A., *op.cit.*, p. 319.

⁷⁶⁸ RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», *cit.*, p. 454.

Ello sin perjuicio de que, además, podría ser contraproducente prohibir la utilización de ciertas categorías de datos en el procesamiento en la medida en que su uso permite determinar hasta qué punto dichas categorías condicionan el resultado del algoritmo y qué otros datos se hallan vinculados con la pertenencia a un grupo desaventajado⁷⁶⁹.

En cuarto lugar, en el RGPD tampoco quedan protegidos los perfiles de grupo que se elaboran sin emplear datos personales, sino mediante la identificación de características o atributos de las personas obtenidos a partir de información que está disponible y que pueden ser indicativos de su pertenencia a un grupo determinado. Esto significa que, por ejemplo, si una empresa elabora perfiles basados en patrones generales como los hábitos de consumo en una zona geográfica determinada, pero sin usar datos personales identificables, el RGPD no se aplicaría directamente. En otras palabras, al centrarse en datos personales individuales, no cubre el impacto que los perfiles grupales estereotipados o sesgados pueden tener en decisiones sobre personas concretas.

En quinto lugar, y en la línea de la consideración anterior, hay que tener especialmente en cuenta que el RGPD no se aplica cuando se utilizan datos anonimizados, pero incluso en estos casos es posible extraer información personal a partir de ellos. Éste es el motivo principal por el que un sector doctrinal mantiene que habría que relativizar la importancia de la normativa sobre protección de datos para hacer frente a la discriminación algorítmica, puesto que la misma está dirigida a regular el uso de datos personales y las predicciones algorítmicas funcionan sobre todo con datos anonimizados⁷⁷⁰.

Como explica con claridad HUERGO LORA, para entrenar a un algoritmo que sirva, por ejemplo, para determinar la solvencia crediticia de un potencial prestatario sólo se necesita información sobre la ejecución de préstamos anteriores. Y esta información no es personal puesto que no es preciso conocer el nombre y apellidos de los prestatarios, sino solo aquellas notas de las personas que pudieran ser relevantes, principalmente, los datos relativos a si cumplieron o no el contrato de préstamo: «A partir de esos datos no personales

⁷⁶⁹ Tal como explica SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación: aproximación...», cit., p. 20/31, si no se utilizaran estas categorías sospechosas en el procesamiento, sería más difícil (aunque no imposible) identificar si el algoritmo está infiriendo la pertenencia a un grupo desaventajado de otras categorías de datos. Además, mantiene que «es más sencillo enseñar a los sistemas automatizados a no discriminar con base en las categorías sospechosas si dichos datos se introducen y consideran expresamente. De esta manera se puede enseñar al algoritmo a evitar confundir características (variables) de las que se infiere la pertenencia a un grupo desaventajado con variables que realmente tengan un valor predictivo en la determinación de la variable objetivo».

⁷⁷⁰ En este sentido, HUERGO LORA, A., «Una aproximación a los algoritmos...», cit., p. 53 o ALAMEDA CASTILLO, M.T., *op.cit.*, p. 32.

(anonimizados) se hallarán correlaciones y se elaborarán perfiles que servirán para ayudar a decidir la concesión o denegación de una nueva petición de crédito por parte de un tercero. Por tanto, el peticionario de crédito sí aporta sus datos personales, que son analizados por la entidad financiera mediante un modelo algorítmico pero esa aportación de datos no plantea problemas desde el punto de vista de la normativa de protección de datos, porque es una aportación consciente, consentida y legitimada por ese consentimiento y además por su necesidad para la celebración del contrato (artículo 6 RGPD). Pero el tratamiento de los datos anteriores queda al margen de la normativa de protección de datos porque no son datos personales. Es verdad que, como se ha observado reiteradamente, la anonimización puede fallar en casos concretos, en los que, por el escaso número de personas que se encuentren en una determinada situación, sea fácil identificar al sujeto que reúne esas condiciones. Pero se trata de excepciones y no podemos ignorar que la legislación de protección de datos acepta la anonimización como procedimiento para eliminar el carácter personal de los datos y levantar algunas barreras a su tratamiento».

En sexto lugar, se ha criticado que las posibilidades que brinda el art. 22 RGPD para las decisiones plenamente automatizadas no abarquen las decisiones semiautomatizadas, ya que éstas también pueden basarse, a su vez y de forma determinante, en resultados de un sistema de IA discriminatorio.

Por último, la efectividad de los mecanismos generales de control establecidos en el RGPD (como la elaboración y aprobación de códigos de conducta y mecanismos de certificación o las evaluaciones de impacto) que podrían compensar las deficiencias derivadas de centrar el sistema de protección de datos en la autonomía del interesado, está siendo muy limitada por varias razones: primera, porque la implementación de los códigos de conducta es voluntaria; y segunda, porque aunque sea obligatorio realizar evaluaciones de impacto cuando el tratamiento entrañe un alto riesgo para los derechos y libertades de las personas físicas, las autoridades de protección de datos no disponen de recursos suficientes para controlar que efectivamente se están llevando a cabo⁷⁷¹.

Pues bien, algunas de las críticas que se acaban de exponer respecto a las debilidades del RGPD para luchar contra la discriminación algorítmica, pueden considerarse superadas a día de hoy teniendo en cuenta, por un lado, que tanto el RIA como la reciente doctrina del TJUE sentada en su sentencia de 7 de diciembre de 2023 (analizada *supra*), avalan el derecho a una explicación *ex post* y la necesidad de interpretar ampliamente el art. 22 RGPD extendiendo

⁷⁷¹ Reflejo de esta situación es el caso del algoritmo SyRi ya comentado, donde se puso de manifiesto el incumplimiento de la obligación de realizar evaluaciones de impacto relativas a la protección de datos.

sus garantías también a las decisiones semiautomatizadas; por otro lado, que el RIA se aplica frente a tratamientos de datos personales como no personales y respecto a los *inputs* y los *outputs* del sistema.

A mayor abundamiento, debe destacarse antes de finalizar este apartado, que tanto la Propuesta de Directiva sobre responsabilidad extracontractual en materia de IA, como la ya aprobada Directiva (UE) 2024/2853 sobre responsabilidad por los daños causados por productos defectuosos, incorporan un importante derecho a la exhibición de pruebas y unas presunciones de existencia de defecto, culpa y causalidad, de gran utilidad para poder detectar y probar casos de discriminación algorítmica y soslayar los obstáculos relativos a la opacidad de los sistemas.

2.4. LAS MEDIDAS PREVENTIVAS RECOGIDAS EN EL DERECHO ANTI-DISCRIMINATORIO

Las prohibiciones de discriminación consagradas legalmente se dirigen a evitar la toma de decisiones discriminatorias. En este sentido, operan como mecanismos preventivos del daño. Tal como señala RUBÍ PUIG, estas normas sectoriales, si bien interactúan con el Derecho de daños, establecen sobre todo requisitos que deberán cumplirse antes de la puesta en circulación de los sistemas y otros deberes de precaución posteriores, siendo su objetivo proteger de forma preventiva intereses jurídicamente relevantes⁷⁷².

La Ley 15/2022 integral de igualdad ha adoptado este enfoque *preventivo* del daño que combina con su enfoque *reparador*⁷⁷³. Establece en su art. 1 que su objetivo es garantizar y promover el derecho a la igualdad de trato y no discriminación, así como respetar la igual dignidad de las personas en desarrollo de los arts. 9.2, 10 y 14 CE, y que, para conseguirlo, se regulan derechos y obligaciones de las personas (físicas o jurídicas, públicas o privadas), se establecen principios de actuación de los poderes públicos y se recogen *medidas destinadas a prevenir, eliminar, y corregir toda forma de discriminación en los sectores público y privado*. Posteriormente, el art. 25 señala que la protección frente a la discriminación obliga a *aplicar métodos o instrumentos para su*

⁷⁷² RUBÍ PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 632. Se trata, según este autor, de normas de conducta que van dirigidas a reducir la frecuencia y ocurrencia de daños. No obstante, en opinión de SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación: aproximación...», *cit.*, p. 6/31, teniendo en cuenta que los efectos de dichas prohibiciones tienen lugar como «mecanismos *ex post* de identificación y resarcimiento de vulneraciones al derecho a la no discriminación, deben ser tratados y examinados como instrumentos de reacción frente a las situaciones de discriminación».

⁷⁷³ Vid., Preámbulo, ap. III. *Principios, objetivos, medios y estructura*.

*detección, la adopción de medidas preventivas, y la articulación de medidas adecuadas para el cese de las situaciones discriminatorias*⁷⁷⁴, y que el incumplimiento de dichas obligaciones dará lugar a *responsabilidades administrativas, así como, en su caso, penales y civiles* por los daños y perjuicios que puedan derivarse, y que podrán incluir tanto la restitución como la indemnización, hasta lograr la reparación plena y efectiva para las víctimas. Por tanto, si se produjera una situación de discriminación en el seno de una empresa y no se hubieran cumplido las obligaciones anteriores (esto es, no se hubieran adoptado medidas suficientes para detectar, prevenir y acabar con la discriminación), la empresa podría responder en un triple ámbito: administrativo, penal y civil⁷⁷⁵.

Por lo que se refiere en concreto a la IA y a los mecanismos de toma de decisión automatizados, el art. 23 establece que las administraciones públicas favorecerán la puesta en marcha de mecanismos para que los algoritmos involucrados en la toma de decisiones que se utilicen en el sector público tengan en cuenta *criterios de minimización de sesgos, transparencia y rendición de cuentas, siempre que sea factible técnicamente*. En estos mecanismos se incluirán su diseño y datos de entrenamiento y se abordará su potencial impacto discriminatorio a través de la realización de evaluaciones de impacto que determinen el posible sesgo discriminatorio. Además, en el marco de sus competencias, priorizarán la transparencia en el diseño y la implementación de los algoritmos involucrados en procesos de toma de decisiones y la capacidad de interpretación de las decisiones adoptadas por ellos. Asimismo, tanto las administraciones públicas como las empresas promoverán el uso de una IA ética, confiable y respetuosa con los derechos fundamentales, siguiendo especialmente las recomendaciones de la Unión Europea en este sentido. También se prevé, como dispone el último apartado del precepto, la creación de un sello de calidad de los algoritmos.

Si bien es meritorio que se hayan incorporado en una norma con rango de Ley determinadas exigencias sobre el uso de la IA para la toma de decisiones en el ámbito público, no puede pasarse por alto que el precepto tiene un cariz programático y voluntarista en la medida en que se limita a establecer directrices o principios generales de actuación que deben seguir las administraciones para favorecer o promover la adopción de mecanismos que prevengan

⁷⁷⁴ Estas medidas dirigidas a acabar con la situación discriminatoria se incardinan ya en el terreno reactivo y no preventivo. Nos remitimos al Capítulo V.

⁷⁷⁵ Como señala RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 633, se da una interacción entre el Derecho público sancionador y el Derecho de daños, puesto que la internalización del coste esperado de las multas puede servir también para disuadir de llevar a cabo actividades susceptibles de lesionar bienes e intereses jurídicamente protegidos.

la discriminación algorítmica⁷⁷⁶. No obstante, en un futuro próximo deberán articularse medidas más concretas.

Es ciertamente llamativo, además, que se declare que la Administración habrá de favorecer la implementación de estos mecanismos dirigidos a minimizar sesgos, aumentar la transparencia y rendir cuentas, *sólo cuando sea factible técnicamente*. Por ende, podrían obviarse si las circunstancias técnicas impidieran su aplicación o si el propio sistema de IA utilizado fuera intrínseca o estructuralmente opaco. Dicha condición conlleva que se torne imprescindible saber cuándo es posible o imposible técnicamente implementar dichas medidas.

Sea como fuere, a nuestro juicio, parece que con esta previsión se está abriendo expresamente la puerta a que la Administración utilice algoritmos de caja negra, lo que conllevaría que los interesados no pudieran llegar a conocer las razones o motivos de las decisiones o actos administrativos que les afectan, vulnerando derechos tan importantes como la transparencia o la motivación de las resoluciones. Quizá, en lugar de justificar la conculcación de tales derechos en base a razones de carácter técnico, se tendría que haber establecido la obligatoriedad de utilizar algoritmos de caja blanca o explicables para la toma de decisiones públicas con efectos directos en las personas.

Por otro lado, la Ley 15/2022 también contempla la adopción de medidas de acción positiva al objeto de *prevenir, eliminar y corregir* toda forma de discriminación por las causas previstas en la ley. El art. 33 dispone que «*Con el fin de hacer efectivo el derecho constitucional a la igualdad, los poderes públicos adoptarán medidas de acción positiva por razón de las causas establecidas en esta ley e impulsarán políticas de fomento de la igualdad de trato y no discriminación real y efectiva en las relaciones entre particulares; Las empresas podrán asumir la realización de acciones de responsabilidad social consistentes en medidas económicas, comerciales, laborales, asistenciales o de otra naturaleza, destinadas a promover condiciones de igualdad de trato y no discriminación en el seno de las empresas o en su entorno social. En todo caso, se informará a los representantes de los trabajadores de las acciones adoptadas [...]*». En el escenario algorítmico, estas medidas podrían concretarse, por ejemplo, en promover la presencia equilibrada de mujeres y hombres en el diseño de los algoritmos o en los comités que los auditen tras su introducción en el mercado o puesta en funcionamiento, así como en favorecer el diseño de algoritmos que contemplen la perspectiva de género, la discapacidad u otras características protegidas.

⁷⁷⁶ En este sentido, VELASCO RICO, C.L. «Marco regulatorio de los sistemas algorítmicos y de IA...», *cit.*, p. 21 y doctrina citada. Esto implicará, según este autor, una interpretación excesivamente laxa sobre dichas exigencias.

Por último, con el fin de prevenir daños, también se recoge en dicha Ley una tutela institucional que se concreta en la creación de una *Autoridad Independiente para la Igualdad de Trato y la No Discriminación*. Se trata de un organismo independiente, unipersonal, basado fundamentalmente en la *auctoritas* de su titular, dirigido a promover el cumplimiento del derecho antidiscriminatorio y proteger frente a la discriminación⁷⁷⁷. Dicho organismo tendrá que unirse tanto a la Agencia Española de Supervisión de la Inteligencia Artificial como a la AEPD, en la lucha contra la discriminación algorítmica⁷⁷⁸. Queda por ver cómo se coordinarán entre ellos para controlar estos riesgos y hacer frente a los daños que su uso puede generar.

El fomento de la cooperación entre autoridades para tratar asuntos relacionados con el derecho a la igualdad y no discriminación (organismos de igualdad, entidades de protección de datos con experiencia en el campo de las nuevas tecnologías, asociaciones de consumidores, etc.) ya venía defendiéndose por la doctrina alegando que dicha coordinación podía resultar clave para abordar adecuadamente el fenómeno de la discriminación algorítmica⁷⁷⁹.

3. EL PAPEL DE LA ÉTICA Y LA TECNOLOGÍA EN LA PREVENCIÓN DE LA DISCRIMINACIÓN ALGORÍTMICA

Solucionar el problema del sesgo algorítmico requiere adoptar no sólo medidas *jurídicas*, sino también *técnicas*⁷⁸⁰. Además, unas y otras han de estar presididas o inspiradas en *principios y directrices* éticas. De hecho, desde hace

⁷⁷⁷ Vid., sobre este organismo, el Título III de la Ley 15/2022.

⁷⁷⁸ Una de las principales funciones encomendadas a la Agencia Española de Supervisión de la Inteligencia Artificial, que ha sido la primera creada en Europa, es la de «supervisión, y en su caso sanción, de los sistemas de inteligencia artificial con el objeto de *eliminar o reducir los riesgos para la integridad, la intimidad, la igualdad de trato y la no discriminación, en particular entre mujeres y hombres*, y demás derechos fundamentales que pueden verse afectados por el mal uso de los sistemas». Sobre el objeto y fines de esta agencia, *vid.*, el art. 4 Real Decreto 729/2023, de 22 de agosto, por el que se aprueba el Estatuto de la Agencia Española de Supervisión de la Inteligencia Artificial.

⁷⁷⁹ Así lo apuntan, por todos, GERARDS, J./XENIDIS, R., *op.cit.*, p. 111. Sobre el establecimiento de una red de autoridades de control algorítmico habla también SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 123.

⁷⁸⁰ Con carácter general, la doctrina mantiene que para luchar contra la discriminación algorítmica no se deben contraponer las soluciones técnicas y las jurídicas, ya que las primeras requieren para su implantación y generalización del impulso y promoción de las normas jurídicas. En esta línea, entre otros muchos, SÁEZ LARA, C., *op.cit.*; NAVAS NAVARRO, S., *Daños ocasionados por sistemas de inteligencia artificial...*, *cit.*, p. 37; BAROCAS, S./SELBST, A. D., *op.cit.*; HACKER, P., *op.cit.*, pp. 31-34.

ya más de una década se viene trabajando para implantar una regulación de la IA inspirada en principios éticos que consiga, entre otros objetivos, prevenir sus efectos negativos y, en particular, el uso de algoritmos sesgados o discriminatorios⁷⁸¹.

El respeto por los principios éticos en el desarrollo, despliegue y uso de la IA ha sido destacado constantemente por las instituciones europeas a lo largo de sus diferentes documentos y propuestas, en los que se ha venido manteniendo reiteradamente que sin un marco ético, la IA podría socavar valores tan relevantes como la justicia y la igualdad. Tanto es así que las *Directrices éticas para una IA fiable* del Grupo de Expertos en IA apuntan como primera orientación clave para conseguir una IA confiable, la observancia de ciertos principios éticos como el respeto de la autonomía humana, la prevención del daño, la equidad y la explicabilidad. Además, se afirma que la IA fiable es aquella que además de ser legal, es ética y robusta, y se identifican siete requisitos esenciales para lograrla⁷⁸².

A mayor abundamiento, en la *propuesta de Reglamento que acompañaba la Resolución sobre aspectos éticos de la IA, de 20 de octubre de 2020*, el Parlamento precisaba que los valores y principios éticos de la Unión debían incorporarse en el desarrollo de los sistemas inteligentes *desde el diseño*, ya que sólo así podía conseguirse una IA *fiable, éticamente responsable y técnicamente sólida*. Subrayaba que la confianza de los ciudadanos en la IA solo podía basarse en un *marco regulador ético por defecto y ético desde el diseño* que garantizara que toda IA puesta en funcionamiento respetase plenamente los Tratados, la Carta y el Derecho derivado de la Unión. Y por lo que se refiere concretamente al sesgo y a la discriminación algorítmica de los programas informáticos, los algoritmos y los datos, declaraba su ilicitud, manteniendo que tales problemas debían de abordarse regulando los procesos mediante los cuales se diseñan y despliegan⁷⁸³.

Con carácter general, la importancia de los principios y valores éticos radica en que con ellos pueden cubrirse lagunas legislativas. De hecho, éste es el principal motivo por el que, según los expertos, se está regulando desde la ética en el campo de la IA. Tal como señala IDOIA SALAZAR, teniendo en cuenta

⁷⁸¹ Sobre este problema y los mecanismos que se pueden establecer para evitar resultados discriminatorios en algoritmos basados en *machine learning*, puede verse el interesante Informe del World Economic Forum, *How to Prevent Discriminatory Outcomes*. Accesible en https://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf

⁷⁸² *Directrices éticas*, cit., p. 17. Para este Grupo Independiente de Expertos de Alto Nivel sobre IA, la IA *fiable* es aquella que cumple con todas las leyes aplicables (IA lícita), garantiza el respeto de los principios, valores éticos y derechos fundamentales (IA ética) y es robusta tanto técnica como socialmente, ya que los sistemas de IA, incluso si las intenciones son buenas, pueden provocar daños accidentales (IA robusta).

⁷⁸³ Considerando 22.

que la elaboración de las leyes es lenta y la evolución tecnológica muy rápida, debe fomentarse la ética y la responsabilidad humana (y no de las máquinas) para prevenir daños e impactos negativos de la tecnología⁷⁸⁴. Y como sostiene NAVAS NAVARRO, mientras no contemos con normas jurídicas vigentes en la materia, los principios éticos proporcionan un marco de seguridad y fiabilidad para todos los sujetos implicados en su desarrollo y uso⁷⁸⁵.

En definitiva, las exigencias éticas llegan donde la ley no lo hace o se muestra ambigua, buscando salvaguardar principios fundamentales como la dignidad, la libertad, la democracia, la igualdad, la autonomía personal o la justicia. Su propósito es contrarrestar la rigidez de un razonamiento puramente automatizado, promoviendo el desarrollo de algoritmos transparentes y equitativos que garanticen la protección de los derechos de los ciudadanos sin comprometer la eficacia y precisión de los sistemas.

3.1. SOLUCIONES DESDE EL PUNTO DE VISTA DE LA ÉTICA ALGORÍTMICA

Previamente a exponer algunas de las soluciones que se han venido proponiendo desde el punto de vista ético para evitar o mitigar los sesgos discriminatorios del algoritmo, debemos hacer una precisión terminológica respecto al concepto de ética algorítmica y a su diferencia con la ética de la IA en la que se engloba.

Se entiende por *ética algorítmica* aquella rama de la ética aplicada a la tecnología que explora y trata de resolver el conjunto de cuestiones que afloran a raíz de la programación de un algoritmo; por su parte, *la ética de la IA* es un subcampo de la ética aplicada que se centra en los problemas normativos que plantea el desarrollo, despliegue y utilización de la IA. Su preocupación fundamental es identificar de qué modo puede la IA mejorar o despertar inquietudes para la vida de las personas, ya sea en términos de calidad de vida o de autonomía y libertad humana necesarias para una sociedad democrática⁷⁸⁶.

⁷⁸⁴ Mantiene esta autora que, en la medida en que las máquinas no toman decisiones éticas sino decisiones matemáticas, el que tiene que tener ética es el humano, en <https://www.noentendondanada.es/idoia-salazar-la-ia-puede-hacer-mucho-mal/>. En similar sentido, VESTRI, G., *op.cit.*, p. 394, considera que la atención ha de dirigirse a los humanos que preparan la IA puesto que las máquinas no son más que la extensión de la voluntad de las personas y los sesgos de las máquinas son en realidad sesgos de su programador. Por ello, concluye, «la transparencia y la cultura de la ética deben neutralizar las conductas humanas negativas impidiendo la transferencia de estos comportamientos a los códigos fuente».

⁷⁸⁵ NAVAS NAVARRO, S., *Daños ocasionados por sistemas de inteligencia artificial...*, cit., p. 38.

⁷⁸⁶ *Directrices éticas*, p. 11. Sobre la diferencia entre ambas puede verse TORRALBA ROSELLÓ, F. Y OTROS, «Ètica algorítmica i perspectiva de gènere: de l'opacitat a la transparència», *Ars Brevis: anuario de la Càtedra Ramon Llull Blanquerna*, n. 27, 2021, p. 13.

Así, mientras que la ética de la IA sostiene que los sistemas inteligentes no deberían engendrar discriminaciones, la ética algorítmica se pregunta cómo traducir o integrar estos principios generales en el código informático, es decir, trata de codificar la moral. No obstante, ambos enfoques son necesarios para conseguir una IA fiable⁷⁸⁷.

Las medidas que a continuación se proponen guardan relación o se aproximan en mayor medida al enfoque que plantea la ética algorítmica.

3.1.1. Formación y educación de las personas que trabajan con sistemas de IA

El primero de los requisitos que proponía el Grupo de Expertos en IA en sus *Directrices éticas para lograr una IA fiable* era el relativo a la *intervención y supervisión humanas* de los sistemas. Dicha exigencia implica que los seres humanos que trabajan con algoritmos deben ser capaces de tomar decisiones autónomas con conocimiento de causa en relación con los sistemas de IA⁷⁸⁸. Ahora bien, para ello es imprescindible que se les proporcionen los conocimientos y herramientas necesarios para comprender cómo funcionan y para saber interactuar con ellos de manera satisfactoria. Además, siempre que sea posible, se les debe permitir evaluar por sí mismos o cuestionar los resultados del sistema. Esto solo puede lograrse si se lleva a cabo una labor de concienciación y capacitación de todas las personas que están involucradas en la planificación y desarrollo de los algoritmos, o de aquéllas que toman decisiones basadas en sus resultados.

Se debe informar con carácter general de los riesgos éticos y legales que conlleva el uso de algoritmos, y más concretamente, por lo que a nosotros interesa –esto es, identificar y prevenir la discriminación algorítmica–, sobre los patrones estructurales de exclusión y desventaja social existentes, dado que los riesgos de discriminación podrían identificarse con mayor facilidad si tales patrones se conocen.

No sólo hay que formar en esta materia a los expertos que participan en el diseño y desarrollo del algoritmo, sino a todos los profesionales que interactúan o trabajan con ellos (por ejemplo, jueces, empresarios que los implementan en sus empresas con fines concretos, autoridades de supervisión y control, etc.) e incluso a los propios legisladores⁷⁸⁹. Lo que está claro es que

⁷⁸⁷ *Op.cit.loc.cit.*

⁷⁸⁸ P. 19.

⁷⁸⁹ J. CASTAÑEDA, investigadora principal del estudio «Dealing with Gender Bias Issues in Data-Algorithmic Processes: A Social-Statistical Perspective», ya citado, pone de manifiesto que para

debe proporcionarse formación para identificar y comprender cómo funcionan los sistemas de IA y los riesgos discriminatorios vinculados a su uso, así como para diseñar y/o implementar estrategias de prevención o eliminación de sesgos. Dicha formación será importante, además, para que cualquier persona que deba tomar decisiones basadas en resultados algorítmicos, pueda hacer frente de forma eficaz al peligroso sesgo de la automatización⁷⁹⁰.

En las *Directrices éticas para lograr una IA fiable* se apunta la necesidad de fomentar una mentalidad ética y se mantiene que «la IA fiable promueve la participación con conocimiento de causa de todas las partes interesadas. La comunicación, la educación y la formación desempeñan un papel importante, tanto para garantizar la difusión del conocimiento sobre los efectos potenciales de los sistemas de IA como para que la ciudadanía tome conciencia de que puede participar en la configuración del desarrollo social. Esto incluye a todas las partes interesadas, por ejemplo, las implicadas en la fabricación de productos (los diseñadores y desarrolladores), los usuarios (empresas o individuos) y otros grupos afectados (que quizá no adquieran o utilicen un sistema de IA, pero a quienes afectan las decisiones de estos sistemas, así como la sociedad en su conjunto)».

También se considera en ese documento que se debería impulsar la educación básica sobre la IA entre la sociedad para garantizar que la población cuente con los conocimientos esenciales sobre ella. En esta línea se ha propuesto, como buena práctica a implementar en todos los niveles educativos, la

luchar contra el sesgo discriminatorio del algoritmo es preciso que tanto sus diseñadores como todas las personas involucradas en su desarrollo, estén informadas sobre la posibilidad de que surjan sesgos asociados a su lógica. Además, deben conocer las medidas que existen para minimizarlos y aplicarlas para que no se produzcan, porque solo si son conscientes de las discriminaciones que se dan en la sociedad, sabrán identificar cuando sus desarrollos las reproducen. En el Informe *Algorithms and Human Rights*, elaborado por el Comité de Expertos en intermediarios de internet del Consejo de Europa (MSI-NET), p. 43, se destaca que es esencial garantizar que los profesionales de diversos sectores como la tecnología, ingeniería, Derecho, filosofía, ética y comunicación, participen en discusiones y debates sobre IA, así como promover la alfabetización para que el público pueda llegar a comprender la lógica y el funcionamiento de los algoritmos. En particular, se aboga por que las entidades públicas y los Gobiernos tengan acceso a información lo suficientemente completa como para comprender adecuadamente como toman las decisiones estos sistemas profundamente arraigados en la sociedad.

⁷⁹⁰ Sobre las críticas al desconocimiento que los jueces tienen sobre estas herramientas en las que basan sus decisiones, nos remitimos a lo ya comentado sobre el caso *Loomis* en el capítulo I, apartado 2.2.1 y en el capítulo II, apartado 1.2.1.1. Para SORIANO ARNANZ, A./SIMÓ SOLER, E., *op.cit.*, p. 203, la ventaja de que jueces y magistrados estén debidamente formados en perspectiva de género no se limita a que sus pronunciamientos carezcan de prejuicios negativos, sino que cumple una función preventiva y de detección de argumentos estereotipados por parte de los profesionales que intervienen en el juicio (tanto de los empresarios que utilizan estos sistemas, como de los propios programadores que introducen instrucciones, variables y datos).

realización de cursos sobre humanidades digitales, ciencias sociales digitales o ciencias jurídicas digitales, así como talleres sobre cómo prevenir la discriminación algorítmica⁷⁹¹.

Reflejo de la importancia que ha adquirido en los últimos años la educación y formación en competencias de IA es que el RIA ha incorporado un artículo donde expresamente se recoge esta necesidad. Así, el art. 4, titulado «Alfabetización en materia de inteligencia artificial» se establece que «*Los proveedores y responsables del despliegue de sistemas de IA adoptarán medidas para garantizar que, en la mayor medida posible, su personal⁷⁹² y demás personas que se encarguen en su nombre del funcionamiento y la utilización de sistemas de IA tengan un nivel suficiente de alfabetización en materia de IA, teniendo en cuenta sus conocimientos técnicos, su experiencia, su educación y su formación, así como el contexto previsto de uso de los sistemas de IA y las personas o los grupos de personas en que se utilizarán dichos sistemas*»⁷⁹³.

En el Considerando 20 se explica con más detalle que, con el fin de aprovechar las ventajas de estos sistemas y proteger al mismo tiempo los derechos fundamentales, la salud y la seguridad, y de posibilitar el control democrático, la alfabetización en IA debe dotar a proveedores, responsables del despliegue y personas afectadas, de los conceptos necesarios para *tomar decisiones con conocimiento de causa en relación con los sistemas de IA*⁷⁹⁴. Por lo que se refiere

⁷⁹¹ GERARDS, J./XENIDIS, R., *op.cit.*, p. 147. En el Reglamento que acompañaba la *Resolución sobre aspectos éticos de 20 de octubre de 2020*, apartado 39, el Parlamento apoyaba la creación de programas educativos y actividades de sensibilización pública en torno al impacto social, jurídico y ético de la inteligencia artificial. También sobre educación e IA puede verse MORDUCHOWICZ, R., *La inteligencia artificial ¿Necesitamos una nueva educación?*, UNESCO, 2023, recuperado de <https://unesdoc.unesco.org/ark:/48223/pf0000386262>, donde se afirma que es fundamental actualizar los planes de estudio y las metodologías para integrar la IA de manera efectiva, pero que es fundamental que la educación siga priorizando el pensamiento crítico, la creatividad y la ética, sin depender exclusivamente de la tecnología.

⁷⁹² Asimismo, los responsables del despliegue deben garantizar que las personas encargadas de poner en práctica las instrucciones de uso y la supervisión humana establecidas en el presente Reglamento tengan las competencias necesarias, en particular un nivel adecuado de alfabetización, formación y autoridad en materia de IA para desempeñar adecuadamente dichas tareas. Dichas obligaciones deben entenderse sin perjuicio de otras obligaciones que tenga el responsable del despliegue en relación con los sistemas de IA de alto riesgo con arreglo al Derecho nacional o de la Unión (Considerando 91).

⁷⁹³ El art.3, ap. 56, define la «alfabetización en materia de inteligencia artificial» como «*las capacidades, los conocimientos y la comprensión que permiten a los proveedores, responsables del despliegue y demás personas afectadas, teniendo en cuenta sus respectivos derechos y obligaciones en el contexto del presente Reglamento, llevar a cabo un despliegue informado de los sistemas de IA y tomar conciencia de las oportunidades y los riesgos que plantea la inteligencia artificial, así como de los perjuicios que puede causar*».

⁷⁹⁴ Estos conceptos pueden variar en función del contexto pertinente e incluir el entendimiento de la correcta aplicación de los elementos técnicos durante la fase de desarrollo del sistema de IA, las medidas que deben aplicarse durante su uso, las formas adecuadas de interpretar la infor-

concretamente a estas últimas, se les debe dotar de los *conocimientos necesarios para comprender el modo en que las decisiones adoptadas con la ayuda de la IA tendrán repercusiones para ellas*.

Además, se establece que el Comité Europeo de Inteligencia Artificial deberá apoyar a la Comisión para promover las herramientas de alfabetización en IA, la sensibilización pública y la comprensión de los beneficios, riesgos, salvaguardias, derechos y obligaciones en relación con el uso de sistemas de IA⁷⁹⁵, y que tanto la Comisión como los Estados miembros deberán facilitar la elaboración de códigos de conducta voluntarios para promover dicha alfabetización entre las personas que se ocupan del desarrollo, manejo y uso de la IA⁷⁹⁶.

Esta necesidad de adoptar medidas dirigidas a la alfabetización digital también se recoge en el art. 20 del *Convenio Marco del Consejo de Europa sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho*, de 2024, donde se establece que las partes fomentarán y promoverán la alfabetización y las capacidades digitales para todos los segmentos de la población, incluidas las capacidades específicas para los responsables de la identificación, evaluación, prevención y mitigación de los riesgos que plantean los sistemas de IA.

En definitiva, con esta alfabetización en IA se pretende, por un lado, proporcionar a todos los agentes de la cadena de valor de la IA los conocimientos necesarios para garantizar el cumplimiento y la correcta ejecución del RIA, y por otro lado, que tanto la sociedad en general como sobre todo los que se encargan de diseñar, desarrollar, desplegar o usar estos sistemas, tomen conciencia de sus características, forma de funcionamiento, ventajas y potenciales riesgos. Todo ello coadyuvará a consolidar el camino hacia una IA fiable.

3.1.2. Composición heterogénea de los equipos que diseñan y auditan los algoritmos

Se ha puesto también de relieve la necesidad de garantizar la diversidad e inclusión en los equipos que diseñan los algoritmos. En su Comunicación titulada *Generar confianza en la IA centrada en el ser humano*, la Comisión insistió en que la creación de equipos de diseño diversificados y de mecanismos que garantizaran la participación ciudadana en el desarrollo de la IA, podría ayudar a resolver los problemas de sesgo y discriminación algorítmica⁷⁹⁷.

mación de salida del sistema de IA y, en el caso de las personas afectadas, los conocimientos *necesarios para comprender el modo en que las decisiones adoptadas con la ayuda de la IA tendrán repercusiones para ellas*.

⁷⁹⁵ Vid., también art. 66 «Funciones del Comité». Concretamente, el ap. f).

⁷⁹⁶ Vid., art. 95.c).

⁷⁹⁷ COM/2019/168 final, p. 6.

Por su parte, en su *Resolución sobre aspectos éticos de la IA*, el Parlamento Europeo hacía hincapié en la necesidad de contar con diversos equipos de desarrolladores e ingenieros que trabajasen conjuntamente con los principales agentes de la sociedad, para evitar que los sesgos culturales y de género se incluyesen involuntariamente en los algoritmos, sistemas y aplicaciones IA, y recordaba la necesidad de velar porque los intereses de todos los ciudadanos, incluidos los que están marginados o en situación de vulnerabilidad, como las personas con discapacidad, fueran tenidos en cuenta y estuvieran debidamente representados⁷⁹⁸.

Finalmente, el RIA también se ha hecho eco de esta necesidad y establece el deber de animar a los proveedores y responsables del despliegue de todos los sistemas de IA (sean o no de alto riesgo) y de los modelos de IA, a aplicar con carácter voluntario requisitos adicionales relativos, entre otros extremos, a «*la inclusividad y la diversidad en el diseño y el desarrollo de los sistemas de IA, lo que incluye tener en cuenta a las personas vulnerables y la accesibilidad de las personas con discapacidad, la participación de las partes interesadas, con la participación, según proceda, de las partes interesadas pertinentes, como las organizaciones empresariales y de la sociedad civil, el mundo académico, los organismos de investigación, los sindicatos y las organizaciones de protección de los consumidores, en el diseño y el desarrollo de los sistemas de IA, y la diversidad de los equipos de desarrollo, también por lo que respecta a la paridad de género*»⁷⁹⁹.

La heterogeneidad en la composición de estos equipos conlleva, a nuestro modo de ver, dos requerimientos:

Por un lado, que los equipos que diseñan y desarrollan los algoritmos sean *multidisciplinares* y participen expertos de distintas ramas como juristas, informáticos, especialistas en ética y sociólogos. De esta forma se podrían detectar y solucionar buena parte de los errores de diseño y configuración del algoritmo. El problema actual es que estos equipos están formados básicamente por profesionales de la ciencia de datos o la computación que carecen de conocimientos en campos tan complejos como la protección de derechos fundamentales o la ética algorítmica, con lo que crean sistemas que desde su punto de vista están bien hechos pero que pueden incumplir parámetros o mandatos éticos o legales⁸⁰⁰. Para integrar, en concreto, los conceptos y princi-

⁷⁹⁸ Ap. 39 y 38.

⁷⁹⁹ Considerando 165. Vid., también art. 95.d).

⁸⁰⁰ En opinión de RIVAS VALLEJO, P., «Herramientas desde el derecho antidiscriminatorio...», *cit.*, p. 544, los investigadores del aprendizaje automático deben colaborar con científicos sociales y expertos en humanidades, género, medicina, medio ambiente y Derecho, para que su actuación sea realmente eficaz: «Desde el Derecho la actuación preventiva o reactiva debe ampararse en el alcance de la potencialidad del sesgo en el algoritmo y de actuar sobre él, y para la informática

prios del Derecho antidiscriminatorio en los algoritmos, es necesario que previamente sean conocidos, comprendidos y asumidos por aquéllos que crean sus entradas e instrucciones. Es decir, si los científicos de datos entienden en qué consiste, por ejemplo, la discriminación por asociación y saben que está prohibida, podrían llegar a entender el impacto que podrían tener las inferencias de datos sobre el círculo personal próximo a cada sujeto y tratar de evitarlo o corregirlo⁸⁰¹.

Por otro lado, la heterogeneidad implicaría la observancia de la *diversidad* en cuanto al género, raza, etnia, religión, creencias, discapacidad o edad, en cada una de las fases de desarrollo de la IA. Se ha apuntado que la falta de diversidad entre las personas que integran los equipos es uno de los motivos que puede ocasionar que un sistema resulte sesgado⁸⁰².

Por lo que se refiere concretamente al género, es una realidad que en el sector tecnológico la diversidad brilla por su ausencia, no sólo por el hecho de que el número de mujeres sea muy inferior al de hombres, sino porque dentro del colectivo de mujeres no hay prácticamente representación de razas distinta a la caucásica. El hecho de que, hasta fechas muy recientes, la gran mayoría de programadores y desarrolladores de *software* hayan sido hombres occidentales, caucásicos, heterosexuales y de una clase social acomodada, ha conllevado que la tecnología se haya diseñado desde una perspectiva muy concreta⁸⁰³. De hecho, sectores de las tecnologías digitales emergentes, como los de Silicon Valley, se caracterizan por una crisis de diversidad de género y raza, especialmente en los niveles directivos más altos⁸⁰⁴. En consecuencia, es

resulta imprescindible conocer qué pretende modular el derecho y qué barreras son infranqueables o sobre qué aspectos se debería actuar técnicamente, desde luego contando con expertos en impacto de género, que tampoco abundan en el ámbito del Derecho». Cada vez con mayor frecuencia encontramos trabajos sobre sistemas de IA elaborados por equipos multidisciplinares de investigadores como el dirigido por J. CASTAÑEDA «Dealing with Gender Bias Issues in Data-Algorithmic Processes: ASocial-Statistical Perspective», ya citado. Dicho estudio fue promovido por especialistas en diferentes áreas, entre los que había una socióloga, un antropólogo y expertas en género y estadística. La perspectiva de dicho estudio va más allá de la matemática autónoma asociada al algoritmo ya que éste se entiende como un sistema sociotécnico complejo, tal como señala su IP.

⁸⁰¹ RIVAS VALLEJO, P., «Análisis desde el derecho antidiscriminatorio», *cit.*, p. 427.

⁸⁰² Por todos, SORIANO ARNANZ, A., «Decisiones automatizadas: problemas y soluciones jurídicas...», *cit.*, p. 92.

⁸⁰³ MUÑOZ GUTIÉRREZ, C., *op.cit.*, p. 282 y doctrina citada. Expone NAVAS NAVARRO, S., «La perspectiva de género ...», *cit.*, p.7/22, que el 74% de las personas que trabajan en el diseño y programación de sistemas de inteligencia artificial son hombres, y en concreto, hombres blancos o asiáticos lo que se traduce en una desigual distribución de poder.

⁸⁰⁴ Informe de la Relatora Especial sobre las formas contemporáneas de racismo, discriminación racial, xenofobia y formas conexas de intolerancia de Naciones Unidas, *La discriminación racial y las tecnologías digitales emergentes: un análisis de los derechos humanos*, A/HRC/44/57, junio de 2020, párr. 17. Disponible en: <https://undocs.org/es/A/HRC/44/57>.

bastante probable que la tecnología que se desarrolle en campos donde se excluye de forma desproporcionada a las mujeres y minorías raciales, étnicas o de otro tipo, acabe reproduciendo esa desigualdad.

También se viene manteniendo la necesidad de incentivar una mayor presencia femenina en las denominadas áreas o disciplinas STEM (*Science, Technology, Engineering and Mathematics*) ya que esto contribuiría a lograr la ansiada participación equitativa de hombres y mujeres en los equipos de diseño de los algoritmos y a incluir la perspectiva de género desde el inicio del proceso⁸⁰⁵. Debe tenerse en cuenta que por ejemplo en nuestro país, en la actualidad las tasas de mujeres sobre el total de personas matriculadas en grados universitarios STEM no llega al 50% en casi ningún caso, siendo especialmente bajas en Matemáticas (36%), Física (27%), Telecomunicaciones (23%), o Informática (13%); en los posgrados se reproduce el patrón y en la Formación Profesional la brecha es aún más profunda⁸⁰⁶. Y la situación es similar fuera de nuestras fronteras⁸⁰⁷. Por lo que se refiere a sectores como la IA, solo uno de cada cinco profesionales es una mujer⁸⁰⁸. Por ello, los Estados vienen proponiendo la adopción de medidas y prácticas encaminadas a aumentar la diversidad de género en la educación y el empleo en el ámbito de las STEM⁸⁰⁹.

Sobre esto, y a fin de ampliar las oportunidades de participación de las niñas y las mujeres en la ciencia, la tecnología, la ingeniería y las matemáticas –incluidas las disciplinas relacionadas con las TIC– y mejorar su preparación, empleabilidad, desarrollo de las perspectivas de carrera en condiciones de igualdad y su crecimiento profesional, la UNESCO considera que los Estados Miembros deberían asignar fondos públicos a financiar planes con perspecti-

⁸⁰⁵ En los grados y másteres que incorporan la ciencia de datos en su currículo, se observan deficiencias respecto a la equidad e integración de la perspectiva de género en el diseño de los algoritmos. Sobre esto puede verse, ORTEGA BALANZA, M./ PÉREZ MUÑOZ, E., «Titulaciones informáticas: equidad e integración de la perspectiva de género en el diseño de algoritmos», en Rivas Vallejo (dir.), *Discriminación algorítmica en el ámbito laboral*, 2022, pp. 669-692.

⁸⁰⁶ Datos extraídos del informe *Mujeres en STEM: Desde la educación básica hasta la carrera laboral*, ESADE, marzo 2024, cuyos autores son Cobreros, L., Galindo, J. y Raigada, T., recuperado de <file:///C:/Users/ASUS/Downloads/Mujeres-en-STEM-2024-1.pdf>

⁸⁰⁷ Según una encuesta anual de la *Computing Research Association*, en el año 2019, entre los nuevos graduados de doctorado en IA residentes en Estados Unidos, el 45% eran blancos, mientras que el 22,4% eran asiáticos, el 3,2% eran hispanos y el 2,4% eran afroamericanos. De esos porcentajes, se debe poner atención en que las mujeres graduadas de programas de doctorado de IA han representado el 18,3% de todas las personas graduadas en los últimos 10 años. Estos datos ponen en evidencia la falta de diversidad que hay entre quienes están a cargo de esas labores.

⁸⁰⁸ Según señala Naciones Unidas en <https://www.un.org/es/observances/women-and-girls-in-science-day>

⁸⁰⁹ Algunas de las medidas y prácticas adoptadas por los Estados para aumentar la diversidad de género en la educación y empleo en los campos STEM se explican en el estudio realizado por GERARDS, J./ XENIDIS, R., sobre la discriminación algorítmica en Europa, ya citado, pp. 133 y ss.

va de género, velar por que las políticas digitales nacionales incluyan un plan de acción en materia de género y elaborar políticas pertinentes, por ejemplo, sobre la educación laboral, destinadas a apoyar a las niñas y las mujeres para que no queden excluidas de la economía digital impulsada por la IA. También deberían realizarse inversiones especiales para ofrecer programas específicos y un lenguaje que respete la igualdad de género. Además, se debería incentivar la iniciativa empresarial, la participación y el compromiso de las mujeres en todas las etapas del ciclo de vida de los sistemas de IA, ofreciendo y promoviendo incentivos económicos y reguladores, entre otros estímulos y planes de apoyo, así como políticas encaminadas a lograr una participación equilibrada de hombres y mujeres en la investigación en IA en las universidades y la representación de género en los puestos directivos superiores, las juntas directivas y los equipos de investigación de las empresas digitales y de IA, entre otras acciones positivas que recomienda adoptar⁸¹⁰.

Sin perjuicio de lo anterior, a nuestro modo de ver la diversidad en los equipos debería ir mucho más allá del género e incluir también a personas de grupos especialmente frágiles, como las personas mayores o las personas con discapacidad. Como señala el Grupo de Expertos de Alto Nivel en IA de la Comisión, lo ideal sería que la diversidad en los equipos no solo se materializase en términos de género, cultura y edad, sino también de antecedentes profesionales y conjuntos de competencias.

Dando un paso más, la multidisciplinariedad, diversidad e inclusión también debería exigirse en los equipos encargados de auditar y llevar a cabo el seguimiento posterior de estos sistemas. Se ha puesto de relieve que las métricas para evaluar estos sistemas están desarrolladas por las mismas personas que los han creado, con lo que se genera una especie de círculo vicioso donde se mide lo que se cree que hay que medir. Para reducir al máximo el riesgo de discriminación algorítmica, se propone que los equipos encargados del control en el uso de algoritmos estén integrados por personal de distintas disciplinas (especialmente expertos en protección de derechos fundamentales y ética algorítmica) y que además haya una participación equilibrada de hombres y mujeres⁸¹¹.

No cabe duda, en suma, de que la heterogeneidad en sus diversas manifestaciones y fases de desarrollo de la IA ha de ocupar un lugar destacado en-

⁸¹⁰ Vid., la *Recomendación sobre la ética de la IA* de la UNESCO, cit., p. 33.

⁸¹¹ La UNESCO considera que «los procesos de seguimiento y evaluación deberían asegurar una amplia participación de todas las partes interesadas, entre ellas, aunque no exclusivamente, las personas vulnerables o en situación de vulnerabilidad. Se debería garantizar la diversidad social, cultural y de género, con miras a mejorar los procesos de aprendizaje y fortalecer los nexos entre las conclusiones, la adopción de decisiones, la transparencia y la rendición de cuentas sobre los resultados». *Recomendación sobre la ética de la IA*, cit., p. 40.

tre los principios que deben guiar la lucha contra la discriminación algorítmica en tanto en cuanto garantiza que las necesidades, valores e intereses de los distintos grupos de población estén representados en el diseño, implementación y control posterior de los sistemas. Tal como se afirma en el *Informe de la Relatora Especial sobre las formas contemporáneas de racismo, discriminación racial, xenofobia y formas conexas de intolerancia de Naciones Unidas*, «desarrollar tecnologías que encajen dentro de complejas realidades sociales y de los sistemas existentes exige comprender los contextos sociales, jurídicos y éticos, lo que solo puede hacerse incorporando perspectivas diversas y representativas, así como conocimientos técnicos en distintas disciplinas»⁸¹².

A modo de conclusión podría afirmarse que, aunque es cierto que la heterogeneidad no garantiza en todo caso la no discriminación, sí contribuye a tomar en consideración distintas perspectivas y necesidades para el ejercicio de los derechos y libertades a la hora de desarrollar la IA, lo que ayudará a minimizar el sesgo discriminatorio del algoritmo.

No obstante, más allá de todas estas medidas sin duda recomendables, lo cierto y verdad es que el primer dilema ético que plantea el uso de la IA es la conveniencia (o no) de permitir que las decisiones relevantes en la vida de las personas (o con efectos jurídicos o similares según términos del RGPD) sean adoptadas por estos sistemas que, como se ha ido viendo a lo largo de estas páginas, se basan en probabilidades y no en certezas, en muchas ocasiones son opacos e incomprensibles y a menudo discriminan a personas o grupos de forma injusta sin que ni siquiera seamos conscientes de ello.

El debate acerca de qué decisiones podrían delegarse a las máquinas y qué decisiones tendrían que ser adoptadas siempre y necesariamente por humanos, está inconcluso⁸¹³. En este sentido, la UNESCO, en su *Recomendación sobre la ética de la IA*, hace las siguientes consideraciones: «Puede ocurrir que, en algunas ocasiones, los seres humanos decidan depender de los sistemas de IA por razones de eficacia, pero la decisión de ceder el control en contextos limitados seguirá recayendo en los seres humanos, ya que estos pueden recurrir a los sistemas de IA en la adopción de decisiones y en la ejecución de tareas, pero un sistema de IA nunca podrá reemplazar la responsabilidad final de los seres humanos y su obligación de rendir cuentas. Por regla general, las

⁸¹² Informe de la Relatora Especial sobre las formas contemporáneas de racismo, discriminación racial, xenofobia y formas conexas de intolerancia de Naciones Unidas, «La discriminación racial y las tecnologías digitales emergentes: un análisis de los derechos humanos», A/HRC/44/57, junio de 2020, párr. 17. Disponible en: <https://undocs.org/es/A/HRC/44/57>

⁸¹³ Para GINÈS i FABRELLAS, A., *op.cit.*, se deben fijar límites sobre las decisiones que pueden tomar las máquinas y las que no, sobre los datos que pueden extraerse y los que no y sobre las variables que pueden tomarse en cuenta para tomar decisiones.

decisiones de vida o muerte no deberían cederse a los sistemas de IA»; precisamente por ello, los Estados Miembros deberían velar porque siempre sea posible atribuir la responsabilidad ética y jurídica por el uso del sistema IA a personas físicas o jurídicas⁸¹⁴.

3.2. SOLUCIONES QUE PUEDEN ADOPTARSE DESDE EL ÁMBITO TECNOLÓGICO

Junto con las medidas anteriores, se han propuesto otras soluciones desde el punto de vista de la tecnología con el fin de detectar, minimizar y evitar la discriminación algorítmica. Principalmente, estas estrategias van dirigidas, por un lado, a evitar sesgos en los conjuntos de datos utilizados para desarrollar y entrenar el algoritmo, y por otro lado, a crear algoritmos capaces de detectar y mitigar el potencial discriminatorio de otros algoritmos.

Respecto a esta última estrategia, se ha insistido de forma generalizada en la idea de que el *machine learning* es ambivalente en la medida en que los algoritmos pueden causar discriminación, pero también pueden contribuir a evitarla puesto que son capaces de detectar supuestos de discriminación de forma más eficaz y rápida que los humanos⁸¹⁵. De lo que se trataría, según manifiesta la UNESCO, es de asegurar que mediante la IA no se exacerbe la desigualdad existente en el mundo analógico y de aprovechar al máximo el potencial de los sistemas para conseguir una mayor igualdad de género⁸¹⁶.

En el Dictamen del CESE sobre «Inteligencia artificial: anticipar su impacto en el trabajo para garantizar una transición justa», también se pone de manifiesto esta ambivalencia y se explica que, si bien para algunos el uso de algoritmos predictivos en la selección de personal podría reducir las discriminaciones y favorecer la contratación de los «objetivamente» mejores, para otros, los algoritmos de contratación siempre pueden reflejar los sesgos de sus programadores⁸¹⁷.

Aquellos que defienden la primera postura mantienen que el uso de algoritmos transparentes sería preferible a la toma de decisiones humana en la medida en que el acceso a su «caja blanca» permitiría obtener respuestas sobre el proceso de decisión con mayor profundidad que las respuestas humanas. Por ejemplo, en los procesos de selección laboral el responsable de recursos hu-

⁸¹⁴ Vid., aps. 35 y 36.

⁸¹⁵ Ponen de manifiesto esta idea, entre otros muchos autores, GASCÓN MARCÉN, A., *op.cit.*, p. 336; SÁEZ LARA, C., *op.cit.*, p. 45; o ALAMEDA CASTILLO, M.T., *op.cit.*, p. 16.

⁸¹⁶ *Recomendación sobre la ética de la IA*, cit., p. 32.

⁸¹⁷ P. 6.

manos podría ser una persona con sesgos hacia determinados colectivos; dichos sesgos o prejuicios podrían ser incluso inconscientes. En este caso, sería muy complicado *identificar y probar* que el sesgo humano llevó al responsable a adoptar una decisión discriminatoria o estereotipada y no basada en razones objetivas justificativas de la inidoneidad de un candidato para el puesto de trabajo. Sin embargo, un algoritmo de selección de personal diseñado de forma transparente y auditable, que evaluara los antecedentes presentados sin conocer sus datos personales, garantizaría la igualdad de trato en la selección, y podría ser preferible a una evaluación realizada por personas⁸¹⁸.

Para maximizar el potencial de los algoritmos en la detección y prevención de la discriminación es imprescindible adoptar ciertas cautelas: en primer lugar, elaborar estrategias para evitar sesgos en los conjuntos de datos, principalmente en los datos de entrenamiento del algoritmo, que deberían prepararse cuidadosamente, es decir, seleccionarse y limpiarse para eliminar o reducir sesgos; en segundo lugar, habría que informar adecuadamente a los encargados de tomar la decisión acerca de cómo interpretar el resultado arrojado por el algoritmo, así como de la posibilidad de apartarse de él si lo consideran oportuno; además, si aun así el algoritmo discriminara, podrían utilizarse técnicas de auditoría para identificar el origen de la discriminación y poder corregirlo. En este sentido, es esencial que los sistemas de IA permitan técnicamente el registro automático de eventos, mediante archivos de registro, durante toda su vida útil⁸¹⁹. Se ha destacado la gran utilidad que proporciona esta herramienta para detectar el origen de la discriminación algorítmica al permitir registrar información sobre su funcionamiento, esto es, sobre los datos relativos a su movimiento, actuación y decisión. Esto supondría una gran ventaja frente a la toma de decisiones humana, ya que, mientras que el proceso de toma de decisión de una persona no podría registrarse o auditarse de forma plena o absoluta, quedando ocultos los verdaderos motivos que la sustentan en el fuero interno del que la adopta y siendo por ello extraordinariamente difícil detectar el sesgo (consciente o inconsciente), sí que sería posible hacerlo con la toma de decisiones algorítmica en la medida en que hay mecanismos para visualizar, medir, detectar y corregir el sesgo.

Los expertos afirman que hay muchas técnicas para contrarrestar el sesgo discriminatorio de los algoritmos. No obstante, advierten de que el primer paso es detectarlo. Por ejemplo, en un algoritmo de recomendación musical, si la realidad subyacente tiene un sesgo porque hay muchos más artistas masculinos que femeninos en las listas, y en consecuencia, mucha más gente escucha artistas masculinos que femeninos, el algoritmo aprenderá ese pa-

⁸¹⁸ TOLOSA, P./DIBO, C., *op.cit.*, p. 184.

⁸¹⁹ Así lo requiere el RIA para los sistemas de alto riesgo. Vid., arts. 12 y 19.

trón. Para evitar que acabe reproduciéndolo, o incluso exacerbándolo, hay que conseguir que vea el mismo número de mujeres que de hombres y no vea un 90% masculino y un 10% femenino. Para ello, habría que asignar un peso menor a los ejemplos de hombres que a los de mujeres. Por ejemplo, un peso de 0,1 para los hombres y un peso de 100 a las mujeres. Técnicamente se pueden diseñar algoritmos para que las recomendaciones tengan diversidad y esa diversidad podría ser de género o de cualquier otro tipo (racial, cultural, etc.)⁸²⁰.

A la vista de tales manifestaciones, no cabe duda de la capacidad y eficiencia de los sistemas algorítmicos (en particular de los que emplean *machine learning*) para detectar y mitigar la discriminación, tanto en la aplicación de otros algoritmos como para descubrir patrones de exclusión y desventaja estructurales en sectores sociales específicos⁸²¹.

El Parlamento Europeo estima que, con este enfoque, los programas, los algoritmos y los datos podrían convertirse en instrumentos eficaces para luchar contra la discriminación en determinadas situaciones y para lograr la igualdad de derechos y el cambio social positivo⁸²². En la misma línea, el Consejo de Europa insiste en la necesidad de fomentar el uso de aprendizaje automático y los análisis científicos multidisciplinares para combatir la discriminación⁸²³. Y en muchos Estados miembros se ha defendido la idea de que los sistemas de IA pueden contribuir a reducir el riesgo de discriminación e incluso a ayudar a grupos tradicionalmente desfavorecidos como las mujeres, personas de raza negra o personas con discapacidad a conseguir un mejor acceso a ayudas y beneficios sociales, así como a puestos de trabajo y centros educativos⁸²⁴.

⁸²⁰ Entrevista a Nuria Oliver para RTVE: <https://www.rtve.es/playz/20210309/algoritmo-musical-discrimina-a-mujeres/2059220.shtml>

⁸²¹ Así se extrae del estudio de GERARDS, J./XENIDIS, R., *op.cit.*, p. 122. En la p. 123 se ponen algunos ejemplos de algoritmos específicamente desarrollados con el objetivo de rastrear patrones discriminatorios: por ejemplo, un proyecto piloto que se llevó a cabo en los Países Bajos y que desarrolló un algoritmo para rastrear discriminación en los anuncios de trabajo. El algoritmo encontró más de 40.000 elementos textuales insertos en anuncios de trabajo que podían calificarse como discriminatorios. La Inspección de Trabajo podía utilizar los resultados de este proyecto piloto para rastrear más formas de discriminación en el mercado laboral y como base para desarrollar políticas y leyes para contrarrestarla. Además, dicho análisis también permitía a los empleadores eliminar los textos discriminatorios de sus anuncios de trabajo.

⁸²² Reglamento 2020 aspectos éticos (ap. 31).

⁸²³ Vid., la *Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno*.

⁸²⁴ Ejemplos concretos pueden verse en GERARDS, J./XENIDIS, R., *op.cit.*, pp. 122 y ss. En Noruega, por ejemplo, la empresa de tecnología de la información Evry aumentó la contratación de mujeres después de implementar un nuevo proceso de contratación basado en IA. La empresa afirmó que la IA utilizada en dicho proceso garantizaba que las decisiones se basaran en criterios más objetivos respecto de los utilizados en los procesos tradicionales llevados a cabo por personas.

Veamos algunas de las soluciones tecnológicas que se han apuntado para conseguir estos fines.

3.2.1. Estrategias para evitar sesgos en los conjuntos de datos

Tal como mantiene algún autor, aunque el conjunto de datos con los que se entrena un algoritmo refleje la realidad fáctica, esto no exime al aprendizaje automático de la responsabilidad de combatir estereotipos potencialmente dañinos, ya que, si no fuera así, los resultados sesgados podrían aumentar la desviación estadística amplificando la discriminación. En el ejemplo anterior, aun cuando el algoritmo reflejase la realidad al existir efectivamente más artistas masculinos que femeninos, si esta desviación no se corrige provocaría que la mayoría de recomendaciones musicales fueran de cantantes masculinos, al desplazar a las pocas artistas mujeres hacia abajo en los resultados de búsqueda. Por tanto, es importante tener una alternativa que no esté sesgada⁸²⁵.

Se han propuesto diversas estrategias dirigidas a eliminar sesgos en los conjuntos de datos empleados por el algoritmo. Entre ellas puede mencionarse el «etiquetado de género», técnica que consiste en identificar y etiquetar explícitamente datos de género dentro del conjunto de datos utilizados para entrenar un algoritmo. Esto permite a los desarrolladores detectar qué correlaciones encontradas por un algoritmo pueden tener una connotación de género y deben evitarse, y qué correlaciones reflejan una diferenciación justificable. Si el algoritmo se entrena en consecuencia, se podría evitar que los sesgos de género o la discriminación contaminasen sus decisiones. Ese etiquetado también se podría utilizar en relación a otros motivos de discriminación prohibidos. De hecho, tal como mantienen los expertos, puede ser necesario permitir el uso de características personales sensibles en el proceso de desarrollo y entrenamiento de un algoritmo para que sea posible «desinfectar» el modelo⁸²⁶.

Esta propuesta choca con otras dirigidas a eliminar del sistema aquellos datos sensibles como pueden ser el sexo, el género o la raza, para evitar la discriminación. Pero como ya se dijo, incluso cuando el sistema es entrenado para no tener en cuenta motivos protegidos, o cuando en los datos de entrada ni siquiera aparecen tales criterios, los algoritmos pueden inferir la pertenencia de una persona a una categoría determinada en base a otras variables o puntos de datos aparentemente no relacionados⁸²⁷. Además, debe tenerse en

⁸²⁵ ZOU, J. «Removing gender bias from algorithms», recuperado de <https://theconversation.com/removing-gender-bias-from-algorithms-64721>

⁸²⁶ Sobre esto GERARDS, J./XENIDIS, R., *op.cit.*, p. 149 y doctrina citada.

⁸²⁷ Así lo afirman los expertos. Vid., por ejemplo, el estudio dirigido por Castañeda, citado *supra*. Según afirma esta investigadora «es posible que los procesos algorítmicos discriminen por género, incluso cuando están programados para ser ‘ciegos’ a esa variable».

cuenta que no siempre es conveniente eliminar determinados criterios del sistema para evitar sesgos o discriminación, ya que precisamente dicha supresión es la que podría generar un resultado discriminatorio⁸²⁸. Se ha apuntado que en ámbitos concretos, como la medicina o el sistema penitenciario, la presencia de datos desagregados es esencial, pues variables como el sexo o el género son fundamentales para alcanzar resultados acertados⁸²⁹. Hay que tener en cuenta que los datos desagregados permiten un análisis más detallado y granular, lo que puede revelar patrones y tendencias que no serían visibles en datos agregados. En opinión de NAVAS NAVARRO, «el uso explícito de estos indicadores conduce a soluciones más justas en la medida en que pueden apreciarse las diferencias de manera que se puedan proponer soluciones adecuadas, precisas para cada individuo»⁸³⁰.

Otras estrategias destinadas a mejorar la diversidad y calidad en los conjuntos de datos son el aumento de datos o las estrategias de limpieza que pueden ayudar a eliminar el sesgo en el proceso de selección y preparación de los datos⁸³¹. Afirman los expertos que lo deseable es mitigar el sesgo en las fases más tempranas como, por ejemplo, en los datos de entrenamiento. Esto se denomina *mitigación de preprocesamiento* (porque ocurre antes de la creación del modelo) y puede realizarse a través de diversas técnicas como, por ejemplo, la reponderación de ejemplos subrepresentados⁸³². Además, también existen herramientas que ofrecen conjuntos de datos libres de sesgos para el apren-

⁸²⁸ ORWAT, C., *op.cit.*, p. 56.

⁸²⁹ En este sentido, NAVAS NAVARRO, S., «La perspectiva de género ...», *cit.*, p.12/22, para quien, además del hecho de que desagregar los datos con base en estos indicadores es un mandato legal, en no pocos ámbitos la presencia de estos datos es esencial.

⁸³⁰ *Ibidem*. Se plantea esta autora, además, si sería posible automatizar la «justicia» de las decisiones y menciona que se está trabajando intensamente en ello y que algunos de los nuevos métodos en la detección del sesgo algorítmico son, entre otros, el aprendizaje inverso, el re-muestreo o la introducción de datos falsos en la fase de entrenamiento del sistema para ver cómo neutralizarlos.

⁸³¹ La estrategia de «aumento de datos» (*data augmentation*) consiste en la creación y adición de datos sintéticos o modificados al conjunto de datos original para mejorar la representación y diversidad de los datos utilizados para entrenar un modelo de aprendizaje automático. El objetivo es reducir el sesgo y mejorar la generalización del modelo al asegurar que este se exponga a una variedad más amplia de ejemplos durante el entrenamiento. Puede verse sobre este tema, CALVO MARTIN, J., «El aumento sintético de datos», accesible en <https://www.europeanvalley.es/noticias/el-aumento-sintetico-de-datos/>

⁸³² CANTERO ÁLVAREZ, H., *Justicia algorítmica: el sesgo de los algoritmos*, Universidad Politécnica de Madrid, 2021, p. 13. También se pueden implementar algoritmos de mitigación de preprocesamiento para transformar el conjunto de datos de forma que se consiga mayor igualdad en los resultados positivos en la característica protegida para los grupos privilegiados y no privilegiados. Es decir, mediante métodos de ajuste se pretende obtener un conjunto de datos de entrenamiento transformado para poder verificar si fue efectivo eliminar el sesgo usando la misma métrica que se usó para el conjunto de datos de entrenamiento original.

dizaje automático, lo que permite entrenar y probar los modelos utilizando datos seguros⁸³³.

En suma, como afirma RIVAS VALLEJO, si bien existe toda una doctrina especializada en minería de datos discriminatoria que ha conseguido parametrizar matemáticamente las reglas de la discriminación para introducir posteriormente los necesarios factores de corrección en la fórmula, no se ha encontrado por el momento un método seguro y universal que ofrezca al Derecho la orientación necesaria para encaminar las reformas legales oportunas, puesto que cada propuesta se basa en un enfoque distinto que dificulta su convergencia o se centra en una fase de diseño o aplicación del algoritmo⁸³⁴. En cualquier caso, de lo que no cabe duda es de que las estrategias dirigidas a conseguir datos de calidad para entrenar los sistemas (esto es, datos libres de errores, pertinentes, representativos y completos) tienen un gran potencial para combatir sesgos discriminatorios.

3.2.2. Algoritmos para detectar, mitigar y corregir sesgos de otros algoritmos

El estado actual de la investigación ya permite crear algoritmos capaces de detectar sesgos ocultos en los datos de entrenamiento o que aparecen después en el aprendizaje del modelo, e incluso de reducir o mitigar su impacto o corregir las deficiencias de calidad de los datos o del diseño del sistema⁸³⁵.

Estos algoritmos son necesarios sobre todo cuando se trata de detectar sesgos en los sistemas adaptativos que aprenden continuamente de su propia

⁸³³ *Ibidem*.

⁸³⁴ RIVAS VALLEJO, P., «Sesgos de automatización...», *cit.*, pp.50-51.

⁸³⁵ Se han propuesto algoritmos de ordenamiento que apliquen políticas de acción positiva, como FA*IR, DELTR, o SEARCHF AIR. Se trata de técnicas diseñadas para garantizar una clasificación más justa en sistemas de recomendación, búsqueda y selección. Con dichos algoritmos se pretende reducir la discriminación o el sesgo en los *rankings*, asegurando que grupos históricamente desfavorecidos tengan una representación justa en los resultados. En el caso de la iniciativa FA*IR (Fairness-aware Ranking), se estudiaron datos de ofertas de empleo, de reincidencia de presos y *rankings* de admisión a las universidades para detectar patrones de discriminación en directorios digitales que pudieran favorecer o perjudicar a ciertos colectivos. Lo que hace su algoritmo es ajustar el *ranking* para garantizar que haya una representación equitativa de los distintos grupos sociales en las primeras posiciones de la lista. Por ejemplo, en el caso de un buscador de empleo que prioriza a candidatos según experiencia y habilidades, si hay 100 perfiles de hombres y mujeres igualmente cualificados y en los 10 primeros resultados del buscador solo aparecen hombres, FA*IR ajusta el *ranking* para garantizar que haya mujeres con calificaciones similares en posiciones competitivas. Es decir, el algoritmo incorpora un mecanismo de acción positiva para reorganizar los resultados y evitar la discriminación sin afectar la validez del ranking.

experiencia y de los datos que en ellos se van introduciendo, ya que en tales casos se ha afirmado que no es posible detectar el sesgo en la fase de diseño. Con ellos se podría comprobar periódicamente la existencia de sesgos en el proceso de aprendizaje y corregir cualquier sesgo descubierto a lo largo de su ciclo de vida⁸³⁶.

Sería posible incluso combinar varias de las técnicas o estrategias mencionadas para mitigar el sesgo, de forma que un algoritmo específicamente creado para evitar sesgos, utilizara conjuntos de datos que también estén libres de sesgos para entrenar el modelo, y posteriormente dicho modelo se auditara o añadiera una tarea de comprobación de sesgos para verificar periódicamente su presencia o ausencia.

En definitiva, se puede afirmar que a día de hoy es posible crear sistemas de IA para la toma de decisiones que sean conscientes de la equidad. Por tanto, se deben desarrollar algoritmos de aprendizaje imparcial que incorporen directamente la no discriminación y la equidad como un objetivo, minimizando así el impacto de las características o motivos protegidos en las decisiones algorítmicas.

4. REFLEXIÓN FINAL SOBRE LAS MEDIDAS PREVENTIVAS PARA EVITAR LA DISCRIMINACIÓN ALGORÍTMICA

Sin perjuicio de la mayor o menor efectividad de estas y otras herramientas tecnológicas para evitar sesgos, lo que está claro es que la discriminación algorítmica es más un problema social que un problema técnico o de programación, y no puede abordarse adecuadamente sin configurar medidas desde el ámbito legal y ético para acabar con las desigualdades existentes en la sociedad en general, o en el sector de la IA en particular.

Sin perjuicio de lo afirmado, los estudios evidencian que la implementación de estrategias dirigidas tanto a evitar el sesgo en los conjuntos de datos como a la creación de algoritmos específicamente dirigidos a este fin, permiten hacer frente a la discriminación algorítmica siempre que haya voluntad empresarial.

Ahora bien, el establecimiento y generalización de estas estrategias y soluciones tecnológicas requiere la implementación de políticas públicas y re-

⁸³⁶ CANTERO ÁLVAREZ, H., *op.cit.*, p. 64, indica que algunas de estas herramientas permiten comprobar periódicamente si hay sesgos en el proceso automático y mitigar cualquier sesgo descubierto. Ejemplos concretos de algoritmos desarrollados en distintos países con este fin, pueden encontrarse en: <https://www.agenciasinc.es/Noticias/Nuevos-algoritmos-vigilan-el-buen-comportamiento-de-la-inteligencia-artificial>

gulaciones dirigidas a promover sistemas de IA más equitativos a través, básicamente, de la transparencia, las evaluaciones de impacto y las auditorías algorítmicas⁸³⁷. Son necesarias normas y prácticas concretas dirigidas a aumentar la explicabilidad de los algoritmos, a evidenciar y minimizar su impacto en los derechos fundamentales y a evaluarlos de forma continua durante su ciclo de vida para detectar y corregir posibles sesgos a medida que surgen.

Además, como ya se apuntó, es imprescindible que todas las soluciones jurídicas y tecnológicas se inspiren en principios éticos para alcanzar los objetivos de equidad y justicia en la toma de decisiones automatizadas que utilizan IA.

⁸³⁷ Algunas de las técnicas dirigidas a examinar y determinar la posible existencia de sesgos algorítmicos en las soluciones IA y a garantizar la equidad en la implementación del modelo son las conocidas como *Algorithmic Impact Assessment* (Evaluaciones de Impacto Algorítmico, conocidas como AIA, por sus siglas en inglés), cuya finalidad es aportar elementos para detectar, evidenciar y minimizar el impacto social no deseable en sistemas de toma de decisión automática basados en algoritmos. La AEPD explica que se trata de herramientas que permiten evaluar que los componentes de IA cumplen con determinados parámetros de calidad y que son adecuados para la tarea encomendada. Si bien se han propuesto muchas métricas de sesgo y definiciones de equidad, no hay consenso sobre qué definiciones y métricas deberían usarse en la práctica para evaluar y auditar estos sistemas. No obstante, se han desarrollado diferentes métricas orientadas a evaluar la discriminación algorítmica como *Aequitas*, un kit de herramientas de auditoría de sesgos de código abierto desarrollado por el *Center for Data Science and Public Policy*. Vid., *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial...*, cit., p.37.

Capítulo V

Medidas para hacer frente a los problemas que plantea la discriminación algorítmica (II): los daños derivados de la discriminación y su reparación

1. LAS PRINCIPALES PROPUESTAS DE LAS INSTITUCIONES EUROPEAS PARA SUPERAR LOS RETOS QUE LA IA PLANTEA EN EL ÁMBITO DE LA RESPONSABILIDAD CIVIL

1.1. INTRODUCCIÓN

La relevancia de la responsabilidad por daños causados por la IA y la necesidad de resolver los problemas que plantea, ha sido puesta de manifiesto por las instituciones europeas en muchos de sus trabajos e iniciativas de los últimos años. De hecho, desde un primer momento se consideró un tema tan crucial que se le otorgó prioridad a la hora de emprender la tarea regulatoria sobre esta tecnología⁸³⁸.

⁸³⁸ Así se ponía de manifiesto ya en el *Proyecto de Informe con recomendaciones a la Comisión sobre Normas de Derecho Civil relativas a la robótica*, elaborado por el Grupo de Trabajo del Parlamento Europeo y publicado el 31 de mayo de 2016 donde, tras una introducción y unos principios generales se estimaba necesario comenzar por las cuestiones relativas a la responsabilidad civil.

También en la última propuesta normativa de la Comisión Europea sobre responsabilidad civil extracontractual, se hace hincapié en la relevancia del tema. De hecho, su Exposición de Motivos comienza declarando lo siguiente: «En una encuesta representativa realizada en 2020, la cuestión de la responsabilidad civil figuraba entre los tres principales obstáculos a la utilización de la IA por parte de las empresas europeas. Se citó como el obstáculo externo más importante (43 %) en el caso de las empresas que tienen previsto recurrir a la IA, pero que aún no lo han hecho» (p.1).

Bien es sabido que, en líneas generales, la responsabilidad civil implica la obligación de toda persona de reparar los daños y perjuicios que cause a otra, cuando concurren una serie de presupuestos para ello. En el ámbito de las decisiones adoptadas por sistemas de IA, se ha mantenido más concretamente que la *responsabilidad algorítmica* consistiría en «el proceso de asignar responsabilidad por daños cuando la toma de decisiones algorítmicas produce resultados discriminatorios e inequitativos, o en general, lesivos para los derechos de las personas»⁸³⁹.

Los elementos fundamentales para que surja esa obligación de reparar son, además del propio *daño* (que ha de ser real y evaluable económicamente), la existencia de una acción u omisión *humana* que lo cause, la *relación de causalidad* entre ese comportamiento y el daño, y el *criterio de imputación* del resultado dañoso al agente, criterio que tradicionalmente ha sido la culpa o negligencia, pero actualmente también puede ser el riesgo en los casos previstos por legislador. Por tanto, para que prospere una reclamación de responsabilidad civil extracontractual basada en la culpa o negligencia del agente del daño, el demandante deberá probar dichos elementos, esto es, la acción u omisión dañosa (y culpable como regla general), el daño propiamente dicho y la relación de causalidad entre ellos.

Como se ha ido viendo a lo largo de estas páginas, dichos presupuestos clásicos que conforman la responsabilidad civil se complican cuando el daño es ocasionado por sistemas inteligentes debido a sus características intrínsecas. Así, su complejidad, conectividad, opacidad, vulnerabilidad, apertura a actualizaciones y mejoras tras su puesta en el mercado, capacidad de autoaprendizaje y potencial autonomía en la toma de decisiones, y la pluralidad de agentes involucrados en su diseño, desarrollo, implementación y uso, suponen un reto para la eficacia de las normas sobre responsabilidad civil.

En la propuesta de Directiva sobre responsabilidad extracontractual en materia de IA, de 28 de septiembre de 2022, la Comisión hace hincapié en que las actuales normas nacionales sobre responsabilidad civil son inadecuadas para dar respuesta a los daños causados por productos y servicios en los que se recurre a la IA, por cuanto que las características específicas de esta tecnología pueden dificultar o hacer excesivamente costoso para las víctimas *determinar quién es la persona responsable y probar que se cumplen los requisitos para una demanda de responsabilidad civil admisible*⁸⁴⁰; además, se advierte que los pro-

⁸³⁹ GONZÁLEZ VALVERDE, A., *op.cit.*, p. 1235, quien sostiene que la responsabilidad algorítmica podría considerarse también como un *principio*, aquel por el que un sistema algorítmico debe emplear una serie de controles para garantizar que el operador pueda verificar que actúa de acuerdo con sus intenciones, así como identificar y rectificar resultados perjudiciales.

⁸⁴⁰ En la misma línea advertía el Parlamento Europeo, en su Resolución sobre responsabilidad civil de 20 de octubre de 2020, que la opacidad de los sistemas de IA podía hacer extremadamente

cedimientos judiciales para reclamar la reparación del daño podrían ser muy largos y costosos en comparación con los casos en que no interviniera una IA, con lo que las víctimas podrían verse disuadidas de intentar siquiera obtener una indemnización⁸⁴¹.

Para hacer frente a dichos inconvenientes se abogó, en un primer momento, por la articulación de un régimen de responsabilidad *ad hoc* basado en una responsabilidad objetiva o por riesgo, pero más tarde se consideró innecesario elaborar una nueva regulación sobre la base de que el Derecho de daños actual era suficientemente flexible y adaptable para afrontar los retos de la IA y bastaba, por tanto, con realizar ciertas modificaciones o ajustes concretos con el fin de facilitar a los perjudicados la reparación del daño, pero sin necesidad de reformas profundas en el sistema de responsabilidad civil extracontractual⁸⁴².

A día de hoy, han corrido ríos de tinta sobre cómo debería abordarse la responsabilidad civil por daños ocasionados por sistemas inteligentes y han sido muchas las opiniones acerca de cuál de los regímenes existentes en nuestro ordenamiento sería más apropiado aplicar por analogía a este tipo de casos, así como cuál debería ser el fundamento mismo de dicha responsabilidad. De hecho, esta última cuestión ha sido la más controvertida de toda esta materia⁸⁴³. No es nuestra intención analizar con detalle todas y cada una de las

costoso, o incluso imposible, determinar *quién controlaba el riesgo asociado al sistema de IA, o qué código, entrada o datos provocan en última instancia el funcionamiento lesivo, o la identificación de la relación de causalidad entre el daño y el comportamiento que lo causa*, con el resultado de que las víctimas podían quedar sin recibir una indemnización adecuada.

⁸⁴¹ Vid., p.1. Estas preocupaciones también fueron señaladas por el Parlamento Europeo en su *Resolución de 3 de mayo de 2022, sobre la inteligencia artificial en la era digital* (2020/2266 (INI)).

⁸⁴² Frente a la postura que defendía la elaboración de una regulación nueva basada en una responsabilidad objetiva para los daños ocasionados por la IA, mantienen la suficiencia del Derecho de daños para ofrecer respuestas adecuadas a estos casos, entre otros, RUBÍ PUIG, A., «Retos de la inteligencia artificial ...», *cit.*, p. 58 o ATAZ LÓPEZ, J., «Daños causados por las cosas...», *cit.*, p. 46. Considera este sector doctrinal que se podrían aplicar las normas existentes con las modificaciones o adaptaciones necesarias en virtud de las especiales características de los sistemas de IA, sin necesidad de un cambio de paradigma jurídico. Menciona RUBÍ PUIG algunos ejemplos de mecanismos del Derecho de daños que podrían utilizarse para resolver los casos que se plantean en relación con los daños causados por la IA, como por ejemplo, los *criterios de imputación objetiva*, que sirven para asignar responsabilidad en supuestos en los cuales confluyen varias causas a la producción de un daño, la *inversión de la carga de la prueba*, o la aplicación de doctrinas como la de la *responsabilidad cuasi objetiva* o la *solidaridad impropia*, los cuales deberían ponerse en relación con otros instrumentos habituales en nuestra sociedad para gestionar los problemas de causación de daños como la constitución de fondos o la obligación de contratar seguros de responsabilidad civil. Sin perjuicio, además, de la necesaria interacción entre el Derecho de daños y la actividad administrativa de regulación o de homologación y autorización para prevenir accidentes derivados de la puesta en circulación de nuevos productos.

⁸⁴³ Tal como afirma ATIENZA NAVARRO, M.L., «La responsabilidad civil por daños causados por inteligencia artificial...», *cit.*, p. 387, la decisión de si la responsabilidad debe ser objetiva o subje-

opciones barajadas por la doctrina, ni entrar a exponer sus ventajas e inconvenientes⁸⁴⁴, sino más bien explicar cómo parece que la UE ha decidido abordar este asunto finalmente, tras muchos vaivenes y algunos proyectos normativos como el presentado por el Parlamento Europeo en su Resolución de 20 de octubre de 2020, primer texto que específicamente contenía una propuesta de régimen de responsabilidad civil por daños causados por sistemas de IA que se articulaba bajo la forma legislativa de Reglamento.

En dicha iniciativa, tras advertir de los problemas concretos que se planteaban en este ámbito, el Parlamento descartaba que fuera necesario realizar una revisión completa de las normas sobre responsabilidad civil en toda la Unión y abogaba por llevar a cabo ajustes concretos en la legislación vigente y por introducir nuevas disposiciones bien ponderadas y orientadas para hacer frente a los retos de la IA, a fin de evitar la fragmentación normativa y garantizar una armonización en la materia⁸⁴⁵. Se asentaba sobre tres pilares: por un lado, la necesidad de adaptar (no sustituir) el régimen vigente sobre responsabilidad por productos defectuosos a la nueva realidad que presentan los sistemas de IA, el cual debía seguir aplicándose con carácter preferente cuando fuera posible⁸⁴⁶; por otro lado, un enfoque híbrido de responsabilidad basado en el riesgo implícito de estos sistemas, objetivo para los de alto riesgo y subjetivo con inversión de la carga de la prueba de la culpa para los que no lo fueran; y por último, el rechazo a reconocer personalidad jurídica de los robots.

tiva «puede considerarse como una de las piedras angulares del sistema, que incide, también, de lleno en otras cuestiones íntimamente relacionadas, como el aseguramiento de dicha responsabilidad y su suscripción obligatoria o voluntaria».

⁸⁴⁴ Para un estudio pormenorizado de la cuestión, remitimos a los trabajos de ATIENZA NAVARRO, que ha tratado el tema escrupulosamente desde sus inicios. Particularmente puede verse su monografía *Daños causados por inteligencia artificial...*cit.

⁸⁴⁵ A través de esa Resolución, el Parlamento Europeo pedía a la Comisión que adoptase una propuesta de régimen de responsabilidad civil para la IA basado en el artículo 114 TFUE. Dicha Resolución vino precedida por otros documentos importantes en la materia como, por ejemplo, Resolución del propio Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103(INL)), el Informe NTF 2019 y el Informe que acompañaba al Libro Blanco. Estos dos últimos informes partían de que no era necesario crear un nuevo régimen de responsabilidad civil sino adaptar las normas existentes.

⁸⁴⁶ La Comisión Europea ya advertía sobre la necesidad de adaptar la Directiva 85/374/CEE del Consejo, de 25 de julio de 1985, al mundo tecnológico en su Informe de evaluación del año 2018 (*Informe de la Comisión al Parlamento Europeo, al Consejo y al Comité Económico y Social Europeo sobre la aplicación de la Directiva del Consejo relativa a la aproximación de las disposiciones legales, reglamentarias y administrativas de los Estados miembros en materia de responsabilidad por los daños causados por productos defectuosos (85/374/CEE) [COM (2018) 246 final]*)—, donde ponía de manifiesto que los problemas actuales diferían de los que existían en el mundo predominantemente analógico de 1985, debido a la cada vez mayor interconexión, digitalización, autonomía e inteligencia de los productos, y que ante tales retos era necesaria una respuesta coherente y global como se describía en la iniciativa sobre inteligencia artificial.

En la línea de realizar meras (y mínimas) adaptaciones de la normativa vigente se sitúa la Comisión Europea en su última iniciativa en esta materia, esto es, la propuesta de *Directiva para adaptar las normas de responsabilidad extracontractual en la Inteligencia Artificial*, de 28 de septiembre de 2022⁸⁴⁷. Tanto la propuesta normativa del Parlamento como esta última de la Comisión se analizarán en las siguientes páginas.

Asimismo, teniendo en cuenta que la Directiva propuesta por la Comisión forma parte de un paquete de medidas con las que se pretende adoptar un «enfoque holístico» en la regulación de la IA, abordaremos su relación con otras normas que forman parte de la misma política sectorial: en particular, con la Directiva sobre responsabilidad por productos defectuosos, y con el RIA, cuyas normas fueron analizadas anteriormente.

1.2. LA PROPUESTA DE REGLAMENTO ANEXA A LA RESOLUCIÓN DEL PARLAMENTO EUROPEO SOBRE RESPONSABILIDAD CIVIL, DE 20 DE OCTUBRE DE 2020

La importancia de la propuesta de Reglamento contenida en la Resolución del Parlamento Europeo, de 20 de octubre de 2020, con «Recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial», reside en que contenía un verdadero régimen de responsabilidad civil completo y homogéneo en base al que reclamar los daños derivados de todo tipo de actividades en las que se utilizaran sistemas de IA⁸⁴⁸. Por ello, aunque la Comisión no haya seguido finalmente la estela marcada por el Parlamento, consideramos pertinente exponer las líneas principales seguidas en la regulación que se proponía en dicho Reglamento con el

⁸⁴⁷ No así en la Propuesta de Directiva sobre responsabilidad por productos defectuosos de 2022 en la que, como después de verá, la Comisión finalmente no ha optado solo por modificar o adaptar la normativa de 1985, sino por proponer un régimen nuevo, tal como explica la Exposición de Motivos y el Considerando 5, donde literalmente se dice que, «[h]abida cuenta de la amplitud de las modificaciones que serán necesarias y en aras de la claridad y la seguridad jurídica, la Directiva 85/374/CEE debe derogarse y sustituirse por una nueva Directiva».

⁸⁴⁸ No obstante, la opción de articular un régimen homogeneizador para los daños causados por la IA ha sido criticada por la doctrina en base a que la IA es un fenómeno intrínsecamente heterogéneo y cuyo uso se lleva a cabo en prácticamente todas las actividades de la vida. En este sentido, por todos, ATIENZA NAVARRO, M.L., «La responsabilidad civil por daños causados por inteligencia artificial...», *cit.*, p. 398. A su juicio, estos daños no deben contemplarse como un fenómeno unitario, sino que deben buscarse regulaciones diferentes en función de las características de estos sistemas, ya que, de la misma forma que las normas sobre responsabilidad varían en función del ámbito en que nos encontremos (por ejemplo, sanitario, vehículos a motor, etc), también deberían ser distintas cuando estas actividades se realizan a través de sistemas de IA. En la misma línea, RUBÍ PUIG, A., «Responsabilidad civil e Inteligencia artificial. Un examen crítico...», *cit.*, p. 261.

que se pretendía garantizar la seguridad jurídica y unificar la normativa sobre esta materia en Europa, fijando cuestiones clave como los *sujetos responsables*, el *criterio de imputación* de la responsabilidad, los *daños indemnizables* o el *plazo de prescripción de la acción*, entre otras.

Si bien el propio Parlamento reconoce que eran preferibles reglamentaciones sectoriales específicas para la amplia gama de posibles aplicaciones de la IA, también cree necesario dotar a la UE de un marco jurídico horizontal y armonizado, basado en principios comunes, con el fin de garantizar la seguridad jurídica, establecer una igualdad de normas en toda la Unión y proteger eficazmente los valores europeos y los derechos de los ciudadanos.

En este Reglamento se recogían algunas propuestas ya planteadas anteriormente en el Informe NTF de 21 de noviembre de 2019, como, por ejemplo, la distinción del fundamento de la responsabilidad en función del riesgo del sistema, o la imputación de la responsabilidad al operador de la tecnología. Sin embargo, no se llegaron a incorporar las recomendaciones para facilitar a la víctima la carga de la prueba de la causalidad o del propio fallo o defecto del sistema, que por el contrario son la base de las últimas propuestas regulatorias de la Comisión.

1.2.1. Ámbito de aplicación

Según el art.2.1 del Reglamento, el mismo se aplica «en el territorio de la Unión a aquellos casos en que una actividad física o virtual, un dispositivo o un proceso gobernado por un sistema de IA haya causado daños o perjuicios [...]». Por tanto, cubre los daños causados por todo tipo de IA, al margen de que se base sólo en el propio *software* o *programa informático* o esté integrada en un *dispositivo físico*.

De hecho, el art. 3.a) define los sistemas de IA como aquellos «*basados en programas informáticos o incorporados en dispositivos físicos* que muestran un comportamiento que simula la inteligencia, entre otras cosas, mediante la recopilación y el tratamiento de datos, el análisis y la interpretación de su entorno y la actuación, con cierto grado de autonomía, para lograr objetivos específicos». Dicho concepto suscitó dudas entre los autores acerca de si abarcaba los sistemas basados en algoritmos deterministas o sólo los que gozaban de una capacidad de aprendizaje propia del *machine* o *deep learning*⁸⁴⁹.

Por último, el Reglamento dejaba a salvo otros posibles regímenes de responsabilidad (entre ellos, el relativo a la discriminación), con independencia

⁸⁴⁹ ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...*cit., pp. 290-291.

de que se tratase de un régimen comunitario o nacional⁸⁵⁰. Siendo esto así, podía entenderse que los casos de daños ocasionados por decisiones algorítmicas discriminatorias debían resolverse conjugando ambas regulaciones. Es más, tal como establecía el art. 2.2, todo acuerdo entre el operador del sistema y la víctima del daño para eludir o limitar la aplicación de los derechos y obligaciones impuestas por el Reglamento, sería nulo con independencia de que se hubiese celebrado antes o después de la causación del daño.

1.2.2. Régimen de responsabilidad civil

1.2.2.1. Sujetos responsables

Se parte de que la persona que cree el sistema de IA, lo mantenga, lo controle o interfiera en él debe ser responsable del daño o perjuicio que cause la actividad, el dispositivo o el proceso. Así se desprende de conceptos de justicia en materia de responsabilidad civil ampliamente aceptados y de carácter general en virtud de los cuales, «la persona que crea o mantiene un riesgo para el público es responsable si dicho riesgo causa un daño o un perjuicio y, por lo tanto, debe minimizar *ex ante* o indemnizar *ex post* dicho riesgo»⁸⁵¹.

Se acoge la noción de «gestión del riesgo» o un «enfoque basado en el riesgo» y se atribuye responsabilidad al que se halla en una mejor posición para prevenir el daño, lo que a su vez incentiva a adoptar tecnologías más seguras⁸⁵². El que se halla en una mejor posición para prevenir el daño es aquel que ostenta un mayor grado de control sobre la tecnología porque su funcionamiento depende de sus acciones, identificándose este sujeto en la Propuesta como el «operador» de la tecnología⁸⁵³. Es decir, se propone al operador como sujeto civilmente responsable, entendiendo que será todo sujeto que pueda haber ejercido un grado de control sobre un riesgo del sistema de IA. Así, el concepto de «control» sobre el sistema deviene clave para ser considerado responsable, entendiéndose por control «toda acción de un operador que influya en el funcionamiento de un sistema de IA y, por consiguiente, la

⁸⁵⁰ Disponía el art. 2.3 que la aplicación del Reglamento se entendía «*sin perjuicio de cualquier otra demanda en materia de responsabilidad civil derivada de las relaciones contractuales, así como de la normativa sobre responsabilidad por los daños causados por productos defectuosos, la protección de los consumidores, la lucha contra la discriminación y la protección laboral y del medio ambiente, entre el operador y la persona física o jurídica que haya sufrido un daño o perjuicio a causa del sistema de IA y de que se pueda presentar contra el operador de conformidad con el Derecho de la Unión o nacional*».

⁸⁵¹ Considerados 7 y 8, en relación con los sistemas de IA más avanzados.

⁸⁵² RUBÍ PUIG, A., «Retos de la inteligencia artificial ...», *cit.*, p. 59.

⁸⁵³ Dicho término fue propuesto previamente en el Informe NTF 2019.

medida en que el operador expone a terceros a los potenciales riesgos asociados a la operación y al funcionamiento del sistema de IA» (art. 3 g)⁸⁵⁴. Por tanto, el ejercicio del control vendría a ser cualquier tipo de influencia en el comportamiento del sistema que pueda conllevar riesgos⁸⁵⁵.

Ahora bien, en los sistemas de IA es frecuente que el operador no sea una única persona, ya que, además de aquella que lo utiliza, intervienen otras en su diseño, actualización, soporte o supervisión. Es más, cuanto más sofisticado sea el sistema, podrá aumentar el número de operadores con control y por ende con responsabilidad. Por ello, el art. 3 d) establece que operador será tanto el *operador final* como el *inicial* (siempre que la responsabilidad civil de este último no esté ya cubierta por la Directiva 85/374/CEE) y define al primero como «*toda persona física o jurídica que ejerce un grado de control sobre un riesgo asociado a la operación y el funcionamiento del sistema de IA y se beneficia de su funcionamiento*» (ap.e), y al segundo, como «*toda persona física o jurídica que define, de forma continuada, las características de la tecnología y proporciona datos y un servicio de apoyo final de base esencial y, por tanto, ejerce también grado de control sobre un riesgo asociado a la operación y el funcionamiento del sistema de IA*» (ap.f). A mayor abundamiento, en el Considerando 10 se explica que «aunque en general el operador final parece ser la persona que decide “principalmente” el uso del sistema de IA, el operador inicial podría, de hecho, tener un grado de control más elevado sobre los riesgos operativos». En definitiva, se hace recaer la responsabilidad en quien ejerce el control, obtiene beneficios y asume perjuicios.

El operador final («*fronted operator*») abarcaría conceptos como el de propietario, poseedor o usuario, es decir, sería quien opera con el sistema de IA utilizándolo; mientras que el operador inicial («*backend operator*») sería quien, una vez puesto el sistema en circulación, sigue operando con él sin utilizarlo porque proporciona servicios de soporte, actualización o supervisión, aunque normalmente será el diseñador de la tecnología (que define sus características) o el proveedor de servicios en línea⁸⁵⁶. Lo que debe quedar claro es que tanto uno como otro controlan los riesgos asociados al funcionamiento del

⁸⁵⁴ Sigue diciendo el precepto citado, que «esa acción puede afectar al funcionamiento en cualquier fase al determinar la entrada, la salida o resultados o pueden cambiar funciones o procesos específicos dentro del sistema de IA» y que «el grado en que estos aspectos del funcionamiento del sistema de IA están determinados por la acción depende del nivel de influencia que el operador tenga sobre el riesgo relacionado con la operación y el funcionamiento del sistema de IA».

⁸⁵⁵ Señala LLORENS ESPADA, J., *op.cit.*, p. 594, que en la medida en que el elemento central reside en la capacidad de control que el agente tiene sobre el funcionamiento del sistema, las empresas que introducen un sistema IA en su organización se convertirán en operadores finales si mantienen cierto control de los riesgos del funcionamiento del sistema y obtienen un beneficio del mismo.

⁸⁵⁶ ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...cit.*, p. 293.

sistema y se benefician del mismo. Por tanto, ambos podrán ser responsables de los daños que el mismo ocasione: el operador inicial porque es quien puede tener un grado de control más elevado sobre los riesgos operativos, y el operador final porque es quien decide principalmente el uso del sistema⁸⁵⁷.

No obstante, como apunta NAVAS NAVARRO, en el caso de sistemas expertos que toman decisiones y actúan consecuentemente con independencia de la voluntad humana, ninguno de estos operadores ostentará el control sobre los resultados del sistema, ni siquiera, en ocasiones, el que lo diseñó⁸⁵⁸.

En caso de ser varios los que operan, como es habitual con los sistemas de IA, habría que discernir quién tiene mayor grado de control en cada situación para ser considerado responsable. El propio Parlamento incide en la necesidad de mejorar la trazabilidad de los productos con el fin de identificar mejor a los que intervienen en las distintas fases y poder así determinar también el grado de responsabilidad de cada uno en función del nivel de control sobre el riesgo y el funcionamiento del sistema que cada uno de ellos despliegue. Si fueran varios los agentes con control sobre el sistema, se propone la *responsabilidad solidaria* de todos ellos frente a la víctima, si bien con derecho a reclamar en la vía de regreso unos frente a otros de forma proporcional, debiendo fijarse los porcentajes de responsabilidad en función de los respectivos niveles de control que tengan los operadores sobre el riesgo relacionado con la operación y el funcionamiento del sistema⁸⁵⁹.

⁸⁵⁷ ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...cit.*, p. 296.

⁸⁵⁸ NAVAS NAVARRO, S., «Responsabilidad civil del fabricante y tecnología inteligente...», cit., p. 8/16. Para estos casos propone la aplicación de la regla «*harm within de risk*» como una vía de solución más cercana a la realidad.

⁸⁵⁹ Cuestiona LLORENS ESPADA, J., *op.cit.*, p. 591, que pueda imputarse la totalidad del daño al operador final (por ejemplo, a una empresa que implementa el sistema de IA en su ámbito) cuando su grado de control sobre el sistema sea escaso y el daño derive en mayor medida de otros agentes que participan previamente en la cadena. Ello aun cuando la empresa que responda frente a la víctima pudiera repetir posteriormente ante los otros agentes involucrados, ya que la viabilidad de estas demandas es escasa dada la complejidad de los sistemas y el origen de los agentes interactuantes. En opinión de este autor, es excesivo que la empresa se convierta en responsable solidaria frente a cualquier daño que pudiera derivarse de un sistema de alto riesgo, teniendo en cuenta, además, que al ser una responsabilidad objetiva, podría devenir responsable sin que ni siquiera hubiera mediado culpa o negligencia por su parte. En la p. 594 sigue diciendo que, si el sistema de decisión automatizada utilizado en una empresa genera resultados discriminatorios, es indudable que la empresa que lo utiliza no sólo se consideraría que tiene control sobre el mismo, sino que debería resarcir los daños que la decisión discriminatoria hubiera ocasionado. No obstante, debe tenerse en cuenta que dicha empresa (operadora final) puede desconocer en gran medida cómo funciona un algoritmo concreto o la automatización de los datos. Por ello, salvo que se aprecie un acto ilícito en el actuar empresarial y siempre que se hubiera seleccionado un sistema de IA certificado conforme a un sistema similar al de etiquetado voluntario previsto por la Comisión, habría que entender que la empresa ha actuado conforme a lo que era razonable esperar de ella. No obstante, la empresa operadora ten-

No obstante, a fin de determinar la norma que regiría dicha responsabilidad, el art. 11 establecía varias reglas en caso de pluralidad de operadores cuando uno de ellos fuese, además, el productor: si el operador final fuera también productor (por ejemplo, por haber introducido una programación en el *software* o alterado la fuente de datos con la que opera el sistema), la aplicación del Reglamento prevalecería sobre la Directiva 85/347. Si por el contrario, la figura del productor coincidiera con la del operador inicial, respecto a él prevalecería la Directiva. Por último, si solo hubiera un operador y el mismo fuera también productor, la aplicación del Reglamento prevalecería sobre la Directiva de productos defectuosos.

Como apuntaba ÁLVAREZ OLALLA, no aclaraba el Reglamento si la responsabilidad del operador inicial/productor eliminaba la responsabilidad del operador final, o si ambos respondían solidariamente de la indemnización, aunque fuese por títulos diferentes (uno, el operador inicial/productor por aplicación de la Directiva; y el otro, por aplicación del Reglamento). Para esta autora era preferible la primera interpretación, es decir, cuando hubiera dos operadores y el operador inicial fuese a la vez el productor, éste respondería en virtud de la Directiva de productos defectuosos, de ser ésta aplicable, lo que impediría dirigir la acción contra el operador final. De no ser aquella aplicable, por ejemplo, por que el daño no fuese consecuencia de un defecto del sistema de IA sino de un mal uso del mismo, debía aplicarse a ambos operadores la responsabilidad prevista en el Reglamento⁸⁶⁰.

Respecto a las acciones de regreso, el art. 12 establecía que el operador que fuera considerado responsable solidario con otros operadores frente al afectado, y le hubiera indemnizado íntegramente, podría recuperar parte de la indemnización de otros operadores en proporción a su responsabilidad. Por tanto, el que se hubiera hecho cargo de la indemnización podría repetir después frente a los demás sujetos responsables en la medida en que contribuyeran al daño, lo que por otra parte no es más que el funcionamiento habitual de la solidaridad pasiva. En los casos analizados en este trabajo, dicha situación podría darse cuando, por ejemplo, la empresa que hubiera utilizado

dría que demostrar que supervisó el sistema, que reportó sus posibles irregularidades al fabricante y que, en su caso, actualizó el *software* conforme a lo indicado por el productor.

⁸⁶⁰ ÁLVAREZ OLALLA, M.P., «Propuesta de Reglamento en materia de responsabilidad civil por el uso de inteligencia artificial, del Parlamento Europeo, de 20 de octubre de 2020», *CESCO de Derecho de Consumo*, n. 38, 2021, p.9. A la vista de estas reglas, la doctrina mayoritaria mantuvo que si bien la Propuesta de Reglamento establecía la responsabilidad de aquellos que intervienen en la etapa inicial de creación de estos sistemas (productor, diseñador, fabricante, etc.), consideraba preferible que en estos casos su responsabilidad se rigiese por la Directiva 85/347, cuando fuera posible su aplicación, y con las necesarias adaptaciones legislativas para que fuera perfectamente aplicable a estos supuestos jurídicos.

el *software* para la toma de decisiones hubiera indemnizado al perjudicado por los daños que la decisión discriminatoria le hubiera causado (pensemos, por ejemplo, que se le despidió por motivos raciales), a pesar de haberse probado que el daño provenía de un defecto en los datos de entrenamiento o del propio diseño del sistema, en cuyo caso, podría repetirse contra el programador o entrenador del sistema.

1.2.2.2. Criterio de imputación de la responsabilidad

Hasta llegar a la solución recogida en esta Propuesta, fueron varias las posturas adoptadas en el seno de la UE respecto al criterio de imputación de la responsabilidad en el ámbito de la IA y la robótica⁸⁶¹. En un primer momento, el Parlamento acogió la idea de la «gestión de riesgos» contraponiéndola a la responsabilidad objetiva. En este sentido, en el principio 53 de su Resolución de 2017, consideraba que la Comisión debía evaluar en profundidad si debía aplicarse el enfoque de la responsabilidad objetiva o el de gestión de riesgos a la hora de decidir qué reglas aplicar en materia de responsabilidad por daños, explicando en el principio 55 que el enfoque de gestión de riesgos «no se centra en la persona que actuó de manera negligente como personalmente responsable, sino en la persona que es capaz en determinadas circunstancias de minimizar los riesgos y gestionar el impacto negativo»⁸⁶².

En opinión de NAVAS NAVARRO, se trata de «un (nuevo) criterio de imputación objetiva que se contrapone, de forma incorrecta, a un fundamento de la responsabilidad, cual es el riesgo creado por el desarrollo de una actividad anormalmente peligrosa. No se encuentran en el mismo plano de la discusión ni tampoco son propiamente alternativas. Es más, podrían, si así lo decidiera el legislador europeo, adoptarse ambos, si bien, debería entonces regularse la posibilidad de ejercitar una acción de regreso del sujeto que compensa a la víctima en relación con otros posibles causantes del daño»⁸⁶³. Para otros autores, como ATAZ LÓPEZ, «más que de un concepto jurídico, se trata de un concepto técnico-económico, aunque con clara conexión con principios jurídicos previos, ya que en ella de lo que se trata es de usar como criterio de atribución de responsabilidad el de la posibilidad de minimizar los riesgos. Es decir: la

⁸⁶¹ Sobre las distintas propuestas y opiniones acerca del criterio de imputación y demás elementos de la responsabilidad civil, puede verse ORTIZ FERNÁNDEZ, M., *La responsabilidad civil derivada de los daños causados por sistemas inteligentes...*, *cit.*, p. 103 y ss.

⁸⁶² Como señala ATAZ LÓPEZ, J., «Daños causados por las cosas...», *cit.*, p. 47, para la doctrina española la responsabilidad objetiva es una vieja conocida, pero la idea de «gestión de riesgos» resulta más novedosa.

⁸⁶³ NAVAS NAVARRO, S., «Sistemas expertos ...», *cit.*, p. 6/16.

gestión de riesgos implicaría una responsabilidad civil por omisión (no haber evitado el daño) que presupone un previo establecimiento del deber de ciertas personas de evitar ciertos daños, establecido tras una imprescindible evaluación previa de los riesgos potenciales y la asignación del control de los mismos a determinadas personas»⁸⁶⁴.

La Propuesta de Reglamento del PE parte de este enfoque del riesgo y en lo concerniente al criterio de imputación lo hace depender del nivel de riesgo que conlleven los sistemas. Así, para los de alto riesgo se prevé una responsabilidad objetiva acompañada de un seguro obligatorio⁸⁶⁵, y para los que no son de alto riesgo una responsabilidad subjetiva donde se presume la culpa del operador. El principal motivo para la adopción de estas reglas era evitar que los perjudicados quedasen sin resarcir por no poder probar la culpa del operador.

Con esta duplicidad de regímenes se pretendía encontrar un equilibrio entre los intereses de todos los implicados (operadores y perjudicados), ya que, optar por un régimen único de responsabilidad absoluta para estas tecnologías podría ser contraproducente porque desincentivaría la inversión en IA, y optar por un régimen de responsabilidad subjetivo en el que las reglas procesales sobre la carga de la prueba no se aliviaran y adaptarían a las complejas características de la IA, podría conllevar que los fabricantes, desarrolladores o titulares de estas tecnologías no adoptaran medidas adecuadas para reducir o evitar daños, ante el casi seguro fracaso de las demandas por responsabilidad civil en su contra por falta de prueba de la culpa o la causalidad.

⁸⁶⁴ ATAZ LÓPEZ, J., «Daños causados por las cosas...», *cit.*, p. 48, citando a DÍAZ ALABART, autora a quien también cita ANGUITA RÍOS, R.M., *op.cit.*, p. 2568, cuando explica este enfoque. Para esta última autora, el término gestión del riesgo «se refiere a las posibles consecuencias indeseadas derivadas del uso de la tecnología cuya aparición no pueden determinarse a priori. Para ello se han de planificar las contingencias, evaluar los riesgos potenciales para seleccionar aquellos que tengan mayor impacto negativo y mayor probabilidad de que lleguen a producirse. En un apartado sobre responsabilidad civil como este aplicado a sistemas cognitivos y autónomos es obligado el estudio de este término técnico-económico que se basa más que en el interés de las posibles víctimas, que pueden verse beneficiadas indirectamente, en el interés de las propias empresas relacionadas con la tecnología. La tecnología puede modificarse sustancialmente durante su desarrollo y, por tanto, pueden existir consecuencias no previstas inicialmente que alteren el desenvolvimiento del producto. La gestión de riesgos persigue la identificación y control de aquellos eventos que tienen la capacidad de provocar un cambio no deseado. El simple conocimiento de los riesgos de una actividad ya supone una ventaja al facilitar un estado de alerta sobre los mismos que disminuye sus consecuencias indeseables en caso de producirse».

⁸⁶⁵ El art. 4.4 disponía que, por un lado, el operador final debía garantizar que las operaciones del sistema de IA estuvieran cubiertas por un seguro de responsabilidad civil adecuado en relación con los importes y el alcance de la indemnización previstos en los artículos 5 y 6, y por otro lado, que el operador inicial garantizaría que sus servicios estuvieran cubiertos por un seguro de responsabilidad empresarial o de responsabilidad civil de productos adecuado en relación con tales importes.

El Parlamento reconocía expresamente que este enfoque basado en el riesgo debía basarse en criterios claros y en una definición adecuada de lo que debía considerarse «alto riesgo». Dicha definición se recogía en el art. 3 c), en virtud del cual, debían calificarse como de *alto riesgo* aquellos sistemas de IA autónomos que tuvieran un potencial significativo de causar daños o perjuicios a una o más personas de manera aleatoria, excediendo de lo que cabría razonablemente esperar. Para valorar la magnitud de dicho potencial debía atenderse a la gravedad del posible daño o perjuicio, al grado de autonomía de la toma de decisiones, a la probabilidad de que el riesgo se materializase y al modo y contexto en que se utilizase el sistema de IA⁸⁶⁶.

Además, el Parlamento consideraba que a la hora de determinar si un sistema de IA era o no de alto riesgo también debía tenerse en cuenta el sector en el que cabía esperar que surgieran riesgos importantes y la naturaleza de las actividades realizadas. Sin embargo, en contra de lo que parecía ser su pretensión, no incluyó un listado anexo (revisable y actualizable por la Comisión) enumerando los sistemas de IA de alto riesgo y los sectores críticos en los que tenían que utilizarse⁸⁶⁷.

⁸⁶⁶ Además, como explicaba el Considerando 13, el nivel de *gravedad* debía determinarse en base a la magnitud del daño potencial resultante del funcionamiento en las personas afectadas (incluido el efecto en los derechos fundamentales), el número de personas afectadas, el valor total del posible perjuicio y el daño a la sociedad en su conjunto. La *probabilidad del daño o perjuicio* debía determinarse sobre la base de factores relevantes como el papel de los cálculos algorítmicos en el proceso de toma de decisiones, la complejidad de la decisión y la reversibilidad de los efectos. Y en última instancia, la *forma de uso* debía depender de factores relevantes como el contexto y el sector en el que operase el sistema de IA, si podía tener efectos jurídicos o fácticos sobre derechos importantes de la persona afectada jurídicamente protegidos y de si tales efectos podían evitarse razonablemente.

⁸⁶⁷ El borrador de la Propuesta sí contenía ese Anexo y calificaba de alto riesgo los siguientes sistemas: 1) Aeronave no tripulada en el sentido del artículo 3, punto 30, del Reglamento (UE) 2018/1139, esto es, para los drones que operan o están diseñados para operar de forma autónoma o para ser pilotados a distancia sin un piloto a bordo (en el sector transporte); 2) vehículos con niveles de automatización 4 y 5 con arreglo a la norma SAE J3016 (en el sector transporte); 3) sistemas autónomos de gestión del tráfico (en el sector transporte); 4) robots autónomos (sector asistencial); 5) dispositivos autónomos de limpieza de lugares públicos (sector asistencial). Con carácter previo, en su Libro Blanco sobre IA, la Comisión consideró que la calificación de un sistema como de alto riesgo dependía tanto del sector, como del uso previsto del mismo. Respecto al primer criterio, señalaba que debía considerarse de alto riesgo la aplicación de IA que se empleara en un sector en el que, por las características o actividades que se llevan a cabo normalmente, es previsible que existan riesgos significativos (por ejemplo, la sanidad, el transporte, la energía y determinados ámbitos del sector público). Respecto al segundo criterio, que debía adicionarse al primero, señalaba que la aplicación de IA debía usarse de manera que pudieran surgir riesgos significativos. Por tanto, no toda utilización de la IA en los sectores señalados implicaba por sí misma y necesariamente riesgos significativos (por ejemplo, si bien la atención sanitaria puede ser un sector importante, un fallo en el sistema de asignación de citas de un hospital no supondría en principio un riesgo significativo que justifique la intervención legislativa). La evaluación del nivel de riesgo de un uso determinado podía basarse

1.2.2.3. Causas de exoneración de la responsabilidad

Respecto a las causas de exoneración de responsabilidad, sólo se preveía la fuerza mayor para no responder cuando se tratase de sistemas de alto riesgo. Además, se indicaba expresamente que los operadores tampoco podrían eludir su responsabilidad alegando que los daños fueron causados por el comportamiento autónomo del sistema, previsión que se contempla tanto para los sistemas de alto riesgo como para los que no lo son (arts. 4.3 y 8.2 *in fine*, respectivamente)⁸⁶⁸.

La intervención de un tercero (por ejemplo, un *hacker* que interfiere en el funcionamiento o efectos del sistema) tampoco se preveía como causa de exoneración en los sistemas de alto riesgo, a diferencia de lo que sucedía para los que no lo fueran. En el primer caso, esta norma suponía una excepción a la regla general en materia de responsabilidad civil en la medida en que la interrupción del nexo causal no eximía de responder al operador y se configuraba como una responsabilidad en garantía⁸⁶⁹. En el segundo caso, dicha responsabilidad en garantía o por hecho de otro sólo se aplicaba de forma excepcional y subsidiariamente cuando no fuera posible localizar al tercero o éste fuera insolvente (art. 8.3). Salvo en esos dos casos, cuando el daño hubiera sido causado total o parcialmente por un tercero, la responsabilidad del operador se excluía o aminoraba, respectivamente.

Por lo que se refiere a la culpa exclusiva de la víctima, sólo se preveía como causa de exoneración para los sistemas que no fuesen de alto riesgo (art. 10). No obstante, un sector doctrinal consideraba que también podía exonerar de responsabilidad cuando el sistema fuese de alto riesgo en base al propio art. 10⁸⁷⁰. En este precepto se preveía, en primer lugar, que la culpa ex-

en las repercusiones para las partes afectadas (por ejemplo, el uso de aplicaciones de IA con efectos jurídicos o similares en los derechos de un particular o de una empresa).

⁸⁶⁸ Dicho supuesto, aunque podría calificarse como un caso fortuito, no le eximiría de responsabilidad. Así lo entiende, entre otros, ERCILLA GARCÍA, J., *Normas de Derecho Civil y Robótica. Robots inteligentes, personalidad jurídica, responsabilidad civil y regulación*, Navarra, 2018, p. 87, respecto a los casos de acción autónoma del robot (*culpa in singularitatem*).

⁸⁶⁹ Así la califica ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...* cit., p. 30, destacando la gran trascendencia práctica de esta disposición en este ámbito donde los sistemas son vulnerables a sufrir ciberataques y manipulaciones.

⁸⁷⁰ En este sentido, ORTIZ FERNÁNDEZ, M., *La responsabilidad civil derivada de los daños causados por sistemas inteligentes...*, cit., p. 106. Según ERCILLA GARCÍA, J., *op.cit.*, p. 87, «al tratarse de un sistema de responsabilidad objetiva, la carga de la prueba de la exculpación, ex art. 271 LEC, habrá de recaer sobre los Agentes Intervinientes que habrán de acreditar que el proceder dañoso del robot, no fue debido a un error de programación de su inteligencia artificial (*culpa in codificando*), o a un fallo de fabricación de sus elementos físicos (*culpa in faciendo*), o a una formación en fase de aprendizaje supervisado incorrecta (*culpa in educando*), o de una falta de mantenimiento de los sensores del robot (*culpa in curando*), sino que en todo caso se trataría de culpa exclusiva de la víctima o de fuerza mayor».

clusiva o concurrente de la víctima –o de las personas de las que ésta fuese responsable–, excluiría o reduciría la responsabilidad del operador (art. 10.1); y en segundo lugar, que para probar la negligencia de la persona afectada, el operador podría utilizar los datos generados por el sistema de conformidad con la normativa de protección de datos personales, datos que también podría usar la víctima con fines probatorios o aclaratorios de su demanda. Es más, a fin de facilitar la concreción de responsabilidades en los sistemas de no alto riesgo, el art. 8.4 añadía el *deber del productor de colaborar con el operador o con el afectado* a la hora de aportar información sobre el sistema.

Por lo que se refiere a los *sistemas que no fuesen de alto riesgo*, y habida cuenta de que respecto a ellos se preveía una responsabilidad subjetiva, además de por las circunstancias anteriores (fuerza mayor, intervención de un tercero y culpa exclusiva de la víctima, supuestos todos ellos de interrupción del nexo causal), el operador también podía exonerarse si probaba su actuar diligente. En estos casos, se invertía la carga de la prueba, esto es, se partía de la presunción de culpabilidad del operador y era éste quien debía demostrar su *no culpa* en la causación del daño para no responder.

En cuanto al grado de diligencia exigible, el Considerando 18 explicaba que la diligencia debida que cabía esperar de un operador debía ser acorde con i) la naturaleza del sistema de IA; ii) el derecho jurídicamente protegido potencialmente afectado; iii) el daño o perjuicio potencial que podría causar el sistema de IA; y iv) la probabilidad de dicho perjuicio. Además, se preveía la posibilidad de graduar el nivel de diligencia exigible en función de si el operador era un consumidor o un profesional y se ponía de manifiesto la circunstancia de que, en ocasiones, el operador tenía un conocimiento limitado de los algoritmos y datos utilizados por el sistema, lo que debía tenerse en cuenta a la hora de determinar su responsabilidad.

Respecto a los concretos deberes de diligencia a cumplir, el art. 8.2 establecía una serie de motivos cuya concurrencia debía probar el operador para no responder:

- a) En primer lugar, que el sistema de IA se activó sin su conocimiento, aun habiendo tomado todas las medidas razonables para evitar la activación fuera de su control;
- b) En segundo lugar, que probara haber desplegado la diligencia exigible a través de la realización de unas acciones concretas: 1) Que el sistema que seleccionó era el adecuado para las tareas a realizar (se presumía la adecuación del sistema que hubiese sido certificado por la Comisión⁸⁷¹); 2) Que llevó a cabo una correcta puesta en fun-

⁸⁷¹ GERARDS, J./XENIDIS, R., *op.cit.*, p. 144, ponen de relieve que, a la hora de elegir un sistema de IA, los usuarios tendrían que prestar especial atención a si el modelo a implementar cum-

cionamiento del sistema de IA, acreditando que supervisó real y periódicamente el sistema de IA durante su funcionamiento y que informó al productor de posibles irregularidades; 3) Que llevó a cabo un control de actividades y mantenimiento de la fiabilidad operativa, con instalación de las actualizaciones disponibles.

Surgió la duda entre los autores de si, para acreditar su diligencia, el operador debía probar todas estas acciones o si era suficiente con acreditar solo alguna de ellas. La doctrina mayoritaria defendió, con acierto a nuestro juicio, la primera opción⁸⁷². Además, también se entendió que la prueba de la diligencia debida se reducía a la acreditación de los extremos previstos en el precepto en las condiciones mencionadas y no a cualquier otra circunstancia⁸⁷³.

Por último, respecto a si los *riesgos del desarrollo* debían considerarse como causa de exoneración de responsabilidad del operador en este ámbito, y ante la falta de mención en la Propuesta, un sector doctrinal consideraba que, para los sistemas de no alto riesgo respecto de los que la responsabilidad es subjetiva, habría que atender a las circunstancias del caso concreto para valorar una posible exención de responsabilidad del operador por este motivo. De esta forma, habría supuestos en que los riesgos del desarrollo aconsejarían que no se atribuyera responsabilidad por un determinado hecho, y otros en que debería asumirse el riesgo a pesar del conocimiento limitado del operador. Se trataría de un reparto en la asunción de riesgos ya que en unos casos debería asumirlo el operador y en otros la sociedad, por lo que sería fundamental de-

plía los estándares adecuados o si estaba certificado por agencias externas especializadas, a fin de evitar acusaciones de discriminación. También mantienen que, una medida para gestionar el tema de la responsabilidad, podría ser incluir cláusulas contractuales que les permitieran reclamar daños y perjuicios a las agencias de certificación si se les reclama responsabilidad por daños respecto a un algoritmo certificado.

⁸⁷² En este sentido, por todos, ÁLVAREZ OLALLA, M.P., «Propuesta de Reglamento en materia de responsabilidad civil ...», *cit.*, p. 8 y ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...*cit., p. 310. En sentido contrario, ZORNOZA SOMOLINOS, A., «Breves apuntes a la propuesta de Reglamento del Parlamento Europeo sobre responsabilidad civil en materia de inteligencia artificial», *R.E.D.S.* núm. 17, Julio-Diciembre 2020, p. 99.

⁸⁷³ ÁLVAREZ OLALLA, M.P., «Propuesta de Reglamento en materia de responsabilidad civil ...», *cit.*, p. 8; ATIENZA NAVARRO, M.L., *Daños causados por inteligencia artificial...*cit., p. 311; NAVAS NAVARRO, S., *Daños ocasionados por sistemas de inteligencia artificial...*, *cit.*, p. 66, sostiene que fuera de estos casos previstos en el artículo 8.2 solo dejará de responder por causa de fuerza mayor. Así, no cabe probar el empleo de cualquier diligencia que pudiera ser exigible sino que tendrá que probar las dos causas antes indicadas. A la vista de ello no se trata de un régimen de responsabilidad subjetivo puro, sino más bien de un régimen de responsabilidad cuasi objetivo o de responsabilidad subjetiva objetivizada. También en esta línea, considera ORTIZ FERNÁNDEZ, M., *La responsabilidad civil derivada de los daños causados por sistemas inteligentes...*, *cit.*, p. 111, que de algún modo se objetiva con ello el contenido de la *lex artis* aplicable a este ámbito.

terminar cuál es el riesgo razonable o aceptable que debería ser admitido por los ciudadanos⁸⁷⁴.

1.3. LAS DOS PROPUESTAS DE DIRECTIVAS EN MATERIA DE RESPONSABILIDAD CIVIL DE LA COMISIÓN EUROPEA, DE 28 DE SEPTIEMBRE DE 2022

1.3.1. Su conexión con el RIA

Ni la propuesta inicial de RIA, ni la que finalmente se ha acabado aprobando, recogen normas sobre responsabilidad civil por los daños que los sistemas ocasionen. No obstante, encontramos algunas referencias a esta cuestión en el texto. Así, en el Considerando 79 se reconoce la conveniencia de que «una persona física o jurídica concreta, definida como el proveedor, asuma la responsabilidad asociada a la introducción en el mercado o la puesta en servicio de un sistema de IA de alto riesgo, con independencia de si dicha persona física o jurídica es o no quien diseñó o desarrolló el sistema». Asimismo, dado el complejo desarrollo y distribución de estos sistemas, se introducen disposiciones tendentes a aclarar las obligaciones y responsabilidades de proveedores, responsables del despliegue y otros sujetos de la cadena de valor de la IA y se prevén ciertos supuestos en que los distribuidores, importadores, responsables del despliegue o terceros que suministran al proveedor sistemas, herramientas, servicios, componentes o procesos de IA, deben asumir las obligaciones de éste (art. 25).

Dicha ausencia es intencionada a la vista de que, como explicaba la Comisión en la Exposición de Motivos de su propuesta inicial (ap. 1.3), la misma formaba parte de un *paquete integral* más amplio de medidas dirigidas a abordar los problemas derivados del desarrollo y la utilización de la IA, y se pretendía garantizar su coherencia con otras iniciativas que abordaran los problemas de responsabilidad, las cuales se fundamentarían en la propuesta de RIA y la complementarían, con el objetivo de conseguir una IA fiable.

Se desprende de tales consideraciones que la finalidad del RIA fue, desde un primer momento, asegurar que la IA se utilizase de forma segura a través del establecimiento de medidas preventivas. Es decir, su objetivo no era fijar normas para conseguir la reparación de los daños sino para prevenirlos⁸⁷⁵.

⁸⁷⁴ ORTIZ FERNÁNDEZ, M., *La responsabilidad civil derivada de los daños causados por sistemas inteligentes...*, cit., p.118-120.

⁸⁷⁵ Con acierto señala PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil aplicables a la inteligencia artificial», en N.

En coherencia con lo afirmado, la Comisión Europea publicó el 28 de septiembre de 2022 dos propuestas para regular cuestiones sobre responsabilidad civil bajo la forma de Directivas: una para *adaptar las normas de responsabilidad extracontractual en la Inteligencia Artificial (AI Liability Directive)* –PDRC–, y otra que respondía a la necesidad de *revisar el régimen de responsabilidad aplicable a los daños causados por productos defectuosos* contenido en la Directiva de 1985 –PDPD– para adaptarlo a las exigencias de esta tecnología. La primera de estas Directivas, en la que nos centraremos en este trabajo, reitera su carácter complementario respecto de otras políticas de IA de la Comisión basadas en requisitos preventivos normativos y de supervisión destinados directamente a evitar violaciones de los derechos fundamentales, entre los que menciona expresamente el RIA, el RGPD, la LSD y la legislación de la UE sobre no discriminación e igualdad de trato. Al complementar estas otras vertientes, dicha propuesta pretende proteger el derecho de las víctimas a una indemnización en virtud del Derecho privado, *incluidas las indemnizaciones por violaciones de los derechos fundamentales*.

No cabe duda, por tanto, de que la pretensión de la Comisión era garantizar la coherencia y complementariedad entre el RIA y las Directivas sobre responsabilidad, principalmente la relativa a la responsabilidad civil extracontractual⁸⁷⁶. Para conseguir dicho fin, la Directiva sobre responsabilidad civil remite al RIA en numerosas ocasiones, por ejemplo, a la hora de definir conceptos claves como el de sistema de IA, proveedor o usuario del sistema (art.2 aps. 1, 3 y 4), al mantener la distinción entre IA de alto riesgo y de no alto riesgo, o al reconocer los requisitos de documentación técnica y transparencia exigidos por el RIA haciéndolos operativos en materia de responsabilidad a través de un derecho a la divulgación de información. Al mismo tiempo, incentiva a los proveedores y usuarios de sistemas de IA (*rectius*, «responsables del despliegue» en los términos definitivos del RIA) a cumplir las obligaciones del RIA, ya que su incumplimiento constituirá un elemento importante para que proceda el aligeramiento de la carga de la prueba, como se explicará después.

Álvarez Lata (coord.), *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, 2024, p. 463, que la principal preocupación de la propuesta de RIA es garantizar que las IA que se pongan en funcionamiento en la UE lo hagan con un adecuado nivel de seguridad, pero no está entre sus fines determinar en qué casos y quiénes tienen que responder de los daños producidos.

⁸⁷⁶ Más que de coherencia o complementariedad, la relación entre ambas normas es, en palabras de PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, p. 465, de *dependencia sistemática* de la Directiva respecto al RIA. De hecho, para este autor, no tendría sentido que la Directiva se aprobase y entrase en vigor si no sucede lo mismo antes con el RIA. Por el contrario, la Directiva de productos defectuosos constituye un sistema normativo autónomo que podría regir los casos de responsabilidad por daños causados por una IA que cayese bajo la definición de producto defectuoso al margen de que previamente se hubiese aprobado el RIA.

Por tanto, queda claro que mientras el RIA, entre otras normas, adoptan un enfoque preventivo del daño estableciendo requisitos de seguridad *ex ante*, estas dos últimas propuestas pretenden garantizar la reparación del daño a las víctimas, sin perjuicio de que también desplieguen un efecto preventivo en tanto en cuanto contribuyen a que aquellos que empleen sistemas de IA adopten las medidas de seguridad oportunas para evitar causar daños de los que deban responder. Y es que, como es habitual, el objetivo de las normas sobre responsabilidad es doble⁸⁷⁷: por un lado, garantizar que el afectado reciba una indemnización que repare los daños y perjuicios sufridos, y por otro lado, generar un efecto preventivo o disuasorio sobre las personas físicas o jurídicas que pueden causar daños o perjuicios, buscando la modulación de su comportamiento para evitar pagar una indemnización. A esta función preventiva del Derecho de daños se refiere concretamente la Comisión al afirmar que «[u]nas normas eficaces en materia de responsabilidad también ofrecen un incentivo económico para cumplir las normas de seguridad [las previstas en la Propuesta de Ley IA] y, por lo tanto, contribuyen a evitar que se produzcan daños».

1.3.2. Relación entre ambas

La Comisión pretendía adoptar un enfoque holístico en materia de responsabilidad civil a través de dos vías: la primera, proponiendo *adaptaciones* de la responsabilidad del productor por productos defectuosos en virtud de la Directiva sobre responsabilidad por los daños causados por productos defectuosos, cuya revisión culmina con la PDPD⁸⁷⁸; y la segunda, mediante la adopción de ciertas medidas con las que se pretende una *armonización específica en materia de responsabilidad extracontractual* que se contienen en la PDRC.

Ahora bien, a diferencia del completo régimen de responsabilidad propuesto por el Parlamento en 2020, la armonización pretendida por la Comisión

⁸⁷⁷ Dicho enfoque (según el cual las normas sobre responsabilidad civil tienen una función preventiva del daño y no solo reparadora) está expresamente reconocido en la PDRC (y también lo estaba en la propuesta de Reglamento del Parlamento de 2020) donde se afirma lo siguiente: «Una de las funciones más importantes de las normas de responsabilidad civil es garantizar que las víctimas de daños puedan reclamar una indemnización. Al garantizar una indemnización efectiva, estas normas contribuyen a la protección del derecho a la tutela judicial efectiva y a un juez imparcial (artículo 47 de la Carta de los Derechos Fundamentales de la Unión Europea, denominada en lo sucesivo «la Carta»), al tiempo que incentivan a las personas que, potencialmente, puedan incurrir en responsabilidad civil a prevenir los daños y perjuicios con el fin de evitar que se genere la responsabilidad».

⁸⁷⁸ Téngase en cuenta que, como ya se dijo, en el momento en que se está redactando este trabajo, dicha propuesta se ha convertido en la Directiva (UE) 2024/2853 y, salvo que indiquemos lo contrario, las observaciones realizadas respecto a la propuesta se pueden considerar aplicables también respecto del texto definitivo.

en esta última Propuesta se centra exclusivamente en normas sobre la *carga de la prueba* dirigidas a facilitar a las víctimas la prueba de la culpa y de la relación de causalidad, presupuestos que junto con la acción u omisión ilícita y la existencia del daño configuran la obligación de indemnizar. Es decir, mientras que la PDPD viene a modificar (o más bien sustituir) un régimen de responsabilidad civil ya existente a nivel comunitario con el fin de adaptarlo a la IA, el objetivo de la PDRC es simplemente introducir unas reglas de carácter procesal en los regímenes nacionales de responsabilidad por culpa para mejorar su funcionamiento cuando los daños son causados por una IA, habida cuenta de las dificultades a las que se enfrentan las víctimas para probar tales extremos.

Ambas Propuestas se complementan entre sí para formar un sistema general de responsabilidad civil eficaz adaptado a la era digital y a la IA, con el que se trata de promover la confianza en las tecnologías digitales. Se sigue, así, la línea ya mantenida por el Parlamento en su Resolución sobre responsabilidad civil, de 20 de octubre de 2020, según la cual ambos regímenes (el de responsabilidad civil y el de productos defectuosos), constituían los *dos pilares de un marco común de responsabilidad civil para los sistemas de IA*.

Por lo que se refiere a su ámbito de aplicación, cada una de ellas ampara demandas referidas a distintos tipos de responsabilidad. La PDPD cubre la responsabilidad *objetiva* del productor por defecto del producto. Se aplica a las demandas presentadas por personas físicas⁸⁷⁹ contra el fabricante, cuando el producto defectuoso ha causado muerte o lesiones corporales (incluidos los daños psicológicos reconocidos desde el punto de vista médico), así como daños materiales o destrucción o corrupción de datos que no se utilicen con fines profesionales⁸⁸⁰.

En aras de garantizar la proporcionalidad de la norma, la Comisión precisa que esta Propuesta no va más allá de lo necesario y que por ello no aborda otros tipos de daños, como los derivados de las violaciones de derechos fundamentales como la privacidad o la discriminación, los cuales se tratarían mejor con arreglo a otra legislación. Se deja claro, por tanto, que esta vía no es la más adecuada para reclamar los daños que se ocasionan en los supuestos objeto de estudio en este trabajo. De hecho, el Parlamento, en su Resolución legislativa de 12 de marzo de 2024 que contiene su Posición sobre la PDPD, añadió el Considerando 24 aclarando que «[l]os tipos de daños distintos de los previstos en la presente Directiva, como las pérdidas económicas propiamente dichas,

⁸⁷⁹ Limitación criticada entre otros por GÓMEZ LIGÜERRE, C.I., «La Propuesta de Directiva sobre responsabilidad ...», *cit.*, p. 2, que considera un error que sus normas se restrinjan a los casos en que el producto defectuoso haya causado daños a *personas físicas* porque deja fuera de esa protección a las empresas y entidades que sufran daños a consecuencias de los daños que causen los defectos de los productos que empleen.

⁸⁸⁰ Vid., art. 6 de la Directiva (UE) 2024/2853.

los ataques a la intimidad o la discriminación, no deben dar lugar por sí mismos a la responsabilidad prevista en la presente Directiva. No obstante, la presente Directiva no debe afectar al derecho a indemnización por cualquier daño, incluso moral, en virtud de otros regímenes de responsabilidad».

Por su parte, la PDRC tiene un ámbito de aplicación más amplio que da cobertura a las demandas nacionales de responsabilidad fundadas principalmente en la *culpa de cualquier persona física o jurídica* (proveedores, responsables del despliegue...) con el fin de indemnizar por *cualquier tipo de daño o perjuicio* cubierto por la legislación nacional (incluidos por tanto los derivados de la discriminación o la violación de otros derechos fundamentales como la vida, salud, propiedad, privacidad, etc.) y a *cualquier tipo de víctima* (particulares, empresas, organizaciones, etc.)⁸⁸¹. Además, cubre los daños y perjuicios causados por sistemas de IA tanto si son de alto riesgo como si no⁸⁸².

En base a lo anterior, las demandas por daños derivados de la discriminación algorítmica en principio quedarían cubiertas por la PDRC, puesto que, como mantiene la Comisión Europea, *las nuevas normas establecidas en la PDPD no permiten indemnizar por vulneración de los derechos fundamentales*⁸⁸³. Por tanto, en los casos en que, por ejemplo, alguien fuera rechazado en una entrevista de trabajo debido a un *software* de IA sesgado utilizado para la contratación, la Comisión explica que los perjudicados deberían recurrir a las normas nacionales en materia de responsabilidad para reclamar una indemnización y la PDRC podría ayudarles en sus reclamaciones a través de las medidas que contiene. Es por ello que nos centraremos principalmente en el análisis de dicha propuesta legislativa, sin perjuicio de las oportunas referencias al régimen de responsabilidad por productos defectuosos.

⁸⁸¹ MARTÍN-CASALS, M., «Las propuestas de la Unión Europea para regular la responsabilidad civil ...», *cit.*, p. 70. Para PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, pp. 464-465, el motivo de que el ámbito de aplicación de la PDRC sea más amplio respecto a los daños resarcibles y sujetos responsables que el de la PDPD, se justifica principalmente por su íntima vinculación con los regímenes internos de responsabilidad subjetiva.

⁸⁸² No obstante, como después se explicará, sólo los sistemas de alto riesgo quedarán sometidos a la exhibición de pruebas (ex art.3.1), a diferencia de lo que sucede cuando la demanda se plantea en base al régimen de productos defectuosos que no distingue entre sistemas de alto riesgo o no alto riesgo para aplicar esta medida. Explica ATIENZA NAVARRO, M.L., «La responsabilidad civil por daños causados por inteligencia artificial...», *cit.*, p. 397, que las reglas de la PDRC van destinadas principalmente a sistemas de alto riesgo, mientras que las normas de la PDPD se aplican a los daños causados por cualquier producto (incluidos los sistemas inteligentes tanto de alto riesgo como de no alto riesgo) que quedarán igualmente sometidos a reglas de exhibición de pruebas (art. 8) y a un régimen de presunciones del defecto y/o de la relación de causalidad entre éste y el daño (art. 9).

⁸⁸³ En «Preguntas y respuestas sobre la revisión de la Directiva sobre responsabilidad por los daños causados por productos defectuosos» en https://ec.europa.eu/commission/presscorner/detail/es/ganda_22_5791

Ahora bien, a pesar de este deslinde y de que a juicio de la Comisión no habría superposición entre las reclamaciones presentadas al amparo de cada una de estas Directivas, creemos que el encaje de ambos instrumentos en ocasiones no queda claro y podrían darse casos en que estos regímenes se solaparan⁸⁸⁴.

Sin ir más lejos, los supuestos analizados en este trabajo podrían constituir un buen ejemplo de lo afirmado. Pensemos en que, como apunta con acierto ATIENZA NAVARRO, si bien en la práctica no es frecuente que los sistemas de IA que no están incorporados a bienes corpóreos causen daños personales⁸⁸⁵, tampoco es descartable su causación en todo caso. Este tipo de daños podría ocasionarse, por ejemplo, cuando en base a una decisión discriminatoria del sistema una mujer fuera descartada como donataria de un órgano, ocasionándosele lesiones físicas, e incluso la muerte, como consecuencia de dicha decisión. ¿Podría entonces plantearse la demanda de responsabilidad contra el fabricante del sistema IA en base al régimen de productos defectuosos? A nuestro modo de ver, la respuesta sería afirmativa. Por ello, en aquellos casos excepcionales (pero posibles) en que la discriminación algorítmica llegara a ocasionar daños personales, también podría articularse la demanda contra el fabricante siempre que se dieran los presupuestos de aplicación de la Directiva de productos defectuosos (principalmente, el defecto del sistema). Ahora bien, aun con todo y con eso, en base a dicha normativa únicamente podrían reclamarse los daños por muerte o lesiones corporales (incluidos los daños para la salud psicológica comprobados médicamente), pero no los daños morales autónomos o las pérdidas patrimoniales puras, que quedan excluidos de su ámbito de aplicación. No obstante, estos últimos quedarían cubiertos por la PDRC y podrían reclamarse en base a las normas nacionales,

⁸⁸⁴ Sobre este tema afirma ATIENZA NAVARRO, M.L., «La responsabilidad civil por daños causados por inteligencia artificial...», *cit.*, p. 396, que delimitar el ámbito de aplicación de ambas Directivas es una cuestión más compleja de lo que *a priori* pudiera parecer.

⁸⁸⁵ En el mismo sentido, NAVAS NAVARRO, S., «Responsabilidad civil del fabricante y tecnología inteligente...», *cit.*, mantiene que es difícil que el defecto que presente un software que no está incorporado a un bien corpóreo (*stand-alone-software*) genere el tipo de daños cubiertos por la normativa de productos defectuosos (daños personales, incluida la muerte y daños materiales o a las cosas en determinadas condiciones). En todo caso, podría generar daños económicamente puros como sería el caso de los algoritmos de alta frecuencia que emiten órdenes y contraórdenes al mercado, pero éstos no están comprendidos en el ámbito de aplicación de la Directiva actual, ni tampoco de las normas nacionales de trasposición, habida cuenta de que la primera es una norma de armonización máxima. Otros supuestos de pérdidas económicas puras en el ámbito de los sistemas de IA pueden ser las mencionadas por RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 668: «supuestos de ilícitos relacionados con la personalización de bienes y servicios, en los costes derivados de gestionar individualmente una brecha de seguridad (por ejemplo, cambio de tarjetas de crédito, implantación de sistemas de doble autenticación de los datos, seguimiento de riesgos), o en los supuestos de pérdida o corrupción de datos no personales o sin valor de afección».

beneficiándose las víctimas de las medidas que contiene para aliviar la carga de la prueba de la culpa y la causalidad.

En definitiva, en casos como el descrito (excepcionales pero posibles) se observa que ambas Directivas podrían ser aplicables y habría que analizar el caso concreto para determinar la vía más recomendable para conseguir la reparación. Fuera de tales supuestos, sin embargo, los propios daños indemnizables bajo el régimen de la PDPD harían que careciera de sentido su aplicación a los casos de discriminación algorítmica, ya que no cubre los daños típicos derivados de la discriminación, esto es, los daños morales puros o autónomos⁸⁸⁶ y las pérdidas patrimoniales puras. Sobre esta cuestión se incidirá en el apartado dedicado a los «daños indemnizables».

Por otro lado, también se aclara en la PDRC que la misma cubre las demandas por daños y perjuicios que hayan sido causados por una información de salida –o por la no producción de una información de salida– imputable a un sistema de IA cuando medie *culpa de una persona*, por ejemplo, el proveedor o el responsable del despliegue con arreglo al RIA que no utilizaron datos de calidad o no usaron el sistema conforme a las instrucciones de uso, respectivamente. En otras palabras, cuando el resultado arrojado por el sistema sea consecuencia de una *acción u omisión negligente* de proveedores o responsables del despliegue (lo que además se presumirá cuando se pruebe el incumplimiento de los deberes de conducta recogidos en los arts. 4.3 y 4.4), las demandas para reclamar responsabilidad por daños caerán bajo el paraguas de esta Propuesta.

Por el contrario, cuando los daños hubieran sido causados directamente por una evaluación humana seguida de una acción u omisión humana y el sistema de IA se hubiera limitado a proporcionar información o asesoramiento que después tuvo en cuenta el humano encargado de tomar la decisión final, no se aplicará la Directiva por cuanto estos casos quedan excluidos expresamente de su ámbito de aplicación. Según señala la Comisión en el Considerando 15, la razón de tal exclusión es sencilla porque en tales supuestos «es posible atribuir los daños causados directamente a una acción

⁸⁸⁶ ATIENZA NAVARRO los define como «los que no están relacionados con daños corporales»; MARTÍN-CASALS, M., «Las propuestas de la Unión Europea para regular la responsabilidad civil ...», *cit.*, p. 85, se refiere a ellos como «daño emocional o moral puro, es decir, la lesión a los sentimientos que no está relacionada ni con un daño corporal, ni con la vulneración de un derecho de la personalidad, ni con un daño a las cosas, en su sentido más amplio». En la nota 54 insiste en que los «daños emocionales» (*emotional harm*) o «daños morales puros» son los que se producen con independencia de un daño corporal a la persona del perjudicado o a sus cosas, y que nuestra jurisprudencia suele describir como situaciones de ansiedad, angustia, zozobra, etc., que se dan, por ejemplo, en casos de «falso positivos», por la angustia creada por la diagnosis errónea de una enfermedad grave que en realidad el perjudicado no tenía».

u omisión humana, ya que la información de salida del sistema de IA no se interpone entre la acción humana y el daño, por lo que establecer la causalidad no es más difícil que en situaciones en las que no interviene un sistema de IA».

En consecuencia, cuando se reclamen daños derivados de decisiones semiautomatizadas, esto es, cuando hubiese sido una persona la que hubiese adoptado la decisión final que causa daños (por ejemplo, el empresario que deniega el crédito o la Administración que deniega la asistencia sanitaria, etc.), habiéndose limitado el sistema de IA a hacer una mera recomendación, los demandantes no se podrían beneficiar de las medidas recogidas en la PDRC porque la conducta humana interrumpiría el nexo causal. No obstante, como apunta RUBÍ PUIG, si bien la actuación posterior del sujeto que adopta la decisión podría ser considerada como *causa adecuada del daño* y, en palabras de la Comisión, acreditar la causalidad no sería más difícil que en los casos en que no interviene la IA, no podrían descartarse supuestos de causalidad sucesiva y de incremento del riesgo que sólo podrían acreditarse a partir de la exhibición de pruebas sobre el sistema y del juego de presunciones previstos en la norma europea⁸⁸⁷.

Sin perjuicio de que dichos supuestos no quedan cubiertos por la PDRC y no se les aplican las medidas que prevé, la reclamación de los daños que de ellos deriven podría hacerse en virtud de los regímenes nacionales, y por lo que se refiere a los casos de discriminación algorítmica los demandantes seguirían beneficiándose de la inversión de la carga de la prueba contenida en las normas antidiscriminatorias cuando concurren las condiciones para ello. Además, respecto a las decisiones semiautomatizadas, es imprescindible tener en cuenta tanto la última doctrina sentada por el TJUE en su sentencia de 7 de diciembre de 2023, como el nuevo derecho a la explicación consagrado en el art. 86 RIA, ya analizados en esta obra.

1.3.3. Análisis de la Propuesta de Directiva relativa a la adaptación de las normas de responsabilidad civil extracontractual a la inteligencia artificial

1.3.3.1. Líneas generales y principales novedades

A pesar de que la Comisión Europea afirma que con la PDRC se da continuidad a la Resolución del Parlamento Europeo sobre responsabilidad civil extracontractual, de 20 de octubre de 2020, el régimen jurídico previsto en el

⁸⁸⁷ RUBÍ PUIG, A., «Responsabilidad civil e Inteligencia artificial. Un examen crítico...», *cit.*, pp. 274-275.

Reglamento que la acompañaba poco tiene que ver con el que finalmente se propone en la Directiva⁸⁸⁸.

Para empezar, el instrumento elegido para legislar ha adoptado la forma de Directiva y no de Reglamento, por lo que es fácilmente observable que no se pretende uniformar sino solo armonizar⁸⁸⁹. A mayor abundamiento, y en contra de la que viene siendo tendencia desde hace años en el ámbito europeo, dicha armonización es de mínimos, y no de máximos⁸⁹⁰, con lo cual, «*los Estados miembros podrán adoptar o mantener normas nacionales más favorables para que los demandantes fundamenten sus demandas civiles de responsabilidad extracontractual por daños y perjuicios causados por sistemas de IA, siempre que dichas normas sean compatibles con el Derecho de la Unión*» (art. 1.4). Así pues, como explica el Considerando 14, se podría mantener la inversión de la carga de la prueba en el contexto de regímenes nacionales de responsabilidad subjetiva o los regímenes nacionales de responsabilidad objetiva o sin culpa que puedan resultar de aplicación a los daños causados por sistemas de IA.

Esta disposición es particularmente relevante para los casos de discriminación algorítmica por dos motivos: primero, porque como ya se explicó, permite la aplicación de las normas sobre inversión o reparto de la carga de la prueba previstas en la legislación comunitaria y nacional (medida más favorable que las presunciones refutables recogidas en la PDRC para aligerar al demandante la prueba de la discriminación); y segundo, porque como explicaremos en su momento, permite aplicar la presunción sobre la existencia de daños morales prevista en el art. 27 Ley 15/2022 integral de igualdad, una vez acreditada la discriminación.

⁸⁸⁸ Similares consideraciones hacen MARTÍN-CASALS, M., «Las propuestas de la Unión Europea para regular la responsabilidad civil ...», *cit.*, p. 71, para quien la PDRC «*comporta abandonar, al menos por ahora, el camino señalado por la Propuesta de Resolución del Parlamento Europeo adoptada en octubre de 2020*»; y EVANGELIO LLORCA, R., *op.cit.*, p. 579, para quien la regulación que contiene PDRC «*se aparta totalmente* de la propuesta de Reglamento del PE de 20 octubre de 2020 en materia de responsabilidad civil por el uso de la inteligencia artificial».

⁸⁸⁹ Sostiene RUBÍ PUIG, A., «Responsabilidad civil e Inteligencia artificial. Un examen crítico de la propuesta de Directiva de 28 de septiembre de 2022», en M. Peguera Poch (coord.) *Perspectivas regulatorias de la inteligencia artificial en la Unión Europea*, Madrid, 2023, p. 265, que adoptar una Directiva y no un Reglamento es coherente con la postura que aboga por que los Estados miembros están mejor situados para decidir cómo se debe responder civilmente por los daños causados por la IA debido a sus diferencias culturales sobre la percepción del riesgo o las funciones propias del Derecho de Daños. Sobre todo, teniendo en cuenta que una parte importante de los daños que ocasionen estos sistemas sería compensable a través de un sistema plenamente armonizado como el de la responsabilidad del fabricante por defecto del producto. Sin perjuicio de que, además, las diferencias entre jurisdicciones aportarán evidencias sobre los efectos de aplicar uno u otro régimen.

⁸⁹⁰ A diferencia de la PDPD que adopta un enfoque de armonización máxima (*vid.*, su art. 3).

La Comisión no considera conveniente ni que sean los Estados miembros los que adopten iniciativas particulares para adaptar sus normas a la IA, ni que sean los órganos jurisdiccionales nacionales los que adapten la aplicación de las normas a la tecnología a fin de llegar a resultados justos para las víctimas, ya que esto podría provocar inseguridad jurídica y aumento de costes para las empresas que operan en la UE, así como disparidad jurídica a nivel comunitario⁸⁹¹. Lo que se pretende con la adopción de unas normas armonizadas es precisamente evitar esa fragmentación jurídica resultante de aplicar tantos regímenes de responsabilidad civil como Estados miembros hay en la UE, y acabar con la inseguridad que genera a los empresarios no saber *ex ante* cuál va a ser la normativa aplicable si provocan daños mediante el uso de la IA, inseguridad que aumenta en los casos de comercio transfronterizo donde la incertidumbre abarcaría diferentes jurisdicciones.

No obstante, esa pretendida armonización ha sido muy limitada finalmente. En concreto, ha consistido en la adopción de una serie de normas de carácter procesal relativas a la carga de la prueba que los Estados tendrán que incorporar a sus ordenamientos y que sólo serán de aplicación cuando las demandas se interpongan en el marco de regímenes de responsabilidad subjetiva o por culpa (art. 1)⁸⁹². Sin embargo, no ha entrado a armonizar los aspectos generales y sustantivos de la responsabilidad civil, como la definición de culpa o la causalidad, los diferentes tipos de daños indemnizables, la distribución de la responsabilidad entre varios causantes de los daños, la concurrencia de culpas, el cálculo de los daños y perjuicios o los plazos de prescripción⁸⁹³. Es comprensible, por tanto, que la Comisión mantenga que las medidas previstas en la Directiva pueden encajar sin problemas en los sistemas de responsabilidad civil en vigor, ya que reflejan un enfoque que no depende de la definición de

⁸⁹¹ Apartado primero de la Exposición de Motivos que contextualiza la propuesta.

⁸⁹² No obstante, hay que tener en cuenta que la PDRC adopta un enfoque por fases y que, tal como disponía, tras la primera fase centrada exclusivamente en instaurar medidas específicas relacionadas con la prueba, habría que *reevaluar* la necesidad de adoptar medidas más estrictas o amplias (por ejemplo, instaurar una responsabilidad objetiva en los casos de uso de IA con un perfil de riesgo particular acompañada de un seguro obligatorio –opción por la que se decantaba el Parlamento en 2020–). Se preveía que la revisión se llevara a cabo transcurridos 5 años tras su aprobación, lo que ya no será posible tras su reciente retirada por la Comisión Europea ante la falta de consenso para su adopción.

⁸⁹³ Tal como señala la Comisión, la presente propuesta «no crea ni armoniza los deberes de diligencia ni la responsabilidad civil de las distintas entidades cuya actividad está regulada por dichos actos jurídicos y, por lo tanto, no crea nuevos tipos de demandas de responsabilidad ni afecta a las exenciones de responsabilidad previstas en esos otros actos jurídicos. *La presente propuesta solo introduce aligeramientos de la carga de la prueba para las víctimas de daños causados por sistemas de IA en las demandas que pueden fundamentarse en la legislación nacional o en estas otras leyes de la UE*». Por tanto, a diferencia de la PDPD que combina el enfoque procesal y sustantivo, la PDRC adopta un enfoque básicamente procesal y con un alcance mucho más limitado que las iniciativas que la precedieron.

conceptos como la «culpa» o el «daño», los cuales varían considerablemente entre los Estados miembros⁸⁹⁴. Siendo esto cierto, también lo es que resulta cuanto menos cuestionable, a nuestro juicio, que con sólo un par de medidas procesales dirigidas a aliviar la carga de la prueba de la culpa y la causalidad (que además podrán no aplicarse cuando existan otras normas más favorables para las víctimas en los distintos ordenamientos), se vaya a alcanzar el objetivo perseguido por la Comisión de evitar la fragmentación e inseguridad jurídica sobre el régimen de responsabilidad aplicable, teniendo en cuenta que la opción de los legisladores nacionales para aplicar unas reglas u otras de responsabilidad queda incólume tras la Propuesta. Por ende, los operadores de sistemas de IA podrán quedar sujetos a normas distintas de responsabilidad en función del lugar de producción del daño⁸⁹⁵.

La ausencia de un régimen de responsabilidad más o menos completo y homogéneo como el que propuso el Parlamento Europeo, quizá se deba, en opinión de ATIENZA NAVARRO, a que la Comisión se ha dado cuenta de que no conviene establecer un régimen homogéneo de responsabilidad para los daños causados por sistemas que son heterogéneos, amén de lo difícil que sería este cometido debido a las disparidades existentes en los regímenes de responsabilidad nacionales⁸⁹⁶. Por tanto, no se parte ahora de un régimen de responsabilidad único para todos los Estados miembros sino de la aplicación de los distintos regímenes nacionales para conseguir la reparación de daños causados por la IA.

Si bien cada Estado tendría que adaptar su legislación procesal para trasponer los mecanismos procesales previstos en la PDRC, también es cierto

⁸⁹⁴ En el Considerando 22 se deja claro que no se armonizan las condiciones en las que los órganos jurisdiccionales nacionales establecen la culpa, ni las condiciones relacionadas con los daños, por ejemplo, la cuestión de qué daños son indemnizables.

⁸⁹⁵ Así lo entiende RUBÍ PUIG, A., «Responsabilidad civil e Inteligencia artificial. Un examen crítico...», *cit.*, p. 265, en base al Reglamento Roma II (art. 4.1).

⁸⁹⁶ Sobre la no conveniencia de fórmulas homogéneas o uniformadoras, ATIENZA NAVARRO, M.L., «La responsabilidad civil por daños causados por inteligencia artificial...», *cit.*, p. 398. Para la autora, a la hora de regular esta cuestión deben tenerse en cuenta las peculiaridades y los fines o actividades a que se destinan los sistemas inteligentes, ya que nada tienen que ver los drones con los vehículos autónomos, ni estos con los robots quirúrgicos, por ejemplo. Considera que no hay motivo para dar una regulación única a todas las actividades por el mero hecho de que sean desarrolladas por medio de sistemas de IA. En opinión de RUBÍ PUIG, A., «Responsabilidad civil e Inteligencia artificial. Un examen crítico...», *cit.*, p. 261, la decisión de utilizar un régimen u otro de responsabilidad habría que tomarla después de analizar las consecuencias que se esperan de su aplicación para una determinada actividad, valorando con cuál de ellos se prevendrían en mayor medida los accidentes sin suponer un freno para la innovación tecnológica mediante la imposición de medidas de precaución costosas socialmente. Para este autor, tendría sentido diferenciar distintos supuestos relacionados con los sistemas de IA y someterlos a regímenes de responsabilidad distintos.

que en muchos casos podrían aplicarse los instrumentos ya previstos en los ordenamientos internos para aliviar la carga de la prueba como, en el caso de España, el *principio de disponibilidad y facilidad probatoria* del art. 217.7 LEC, las *presunciones judiciales* del art. 386 LEC o el *principio res ipsa loquitur* («la cosa habla por sí misma»). Además, como se explicó con detalle en el tercer apartado del capítulo tercero, en el Derecho antidiscriminatorio se establece una inversión o traslado de la carga de la prueba cuando el demandante pueda demostrar que hay indicios suficientes para presumir la discriminación. Por eso creemos que, aun cuando esta Propuesta lograba cubrir la laguna que existía en el ámbito de la responsabilidad civil en el sector de la IA, al menos en el Derecho español ya contábamos con principios y mecanismos con los que se podía conseguir el mismo fin pretendido por la Comisión, esto es, suavizar la carga de la prueba en determinados casos y bajo determinadas condiciones. Por ello, consideramos que su reciente retirada pocos días antes de cerrar este trabajo, no tendrá tanto impacto en los casos de discriminación algorítmica que se planteen en nuestro ordenamiento, como en otros ámbitos donde las presunciones de culpa o causalidad hubieran sido de mayor utilidad.

1.3.3.2. *Ámbito de aplicación*

Respecto a su ámbito de aplicación temporal, el art. 1.2 establece su aplicación únicamente a las demandas interpuestas a partir de la fecha de su transposición. Por tanto, hasta dicho momento habrá que solucionar los casos de daños causados por sistemas de IA con las normas nacionales vigentes.

Respecto a su ámbito de aplicación objetivo, el precepto dispone dos cosas: en primer lugar, que la Directiva se aplica a las demandas de responsabilidad civil extracontractual *basadas en la culpa* cuando los daños y perjuicios hayan sido causados por un sistema de IA⁸⁹⁷; y en segundo lugar, que la Directiva no se aplicará a la responsabilidad penal. No obstante, sí afectará a la responsabilidad civil derivada del delito y podrá resultar aplicable a la responsabilidad del Estado en la medida en que las autoridades estatales también están sujetas al RIA y a las obligaciones que impone.

Particularmente relevante es la disposición contenida en el art. 1.3.d), según la cual, la presente Directiva no afectará a «las normas nacionales que determinen

⁸⁹⁷ Como aclara RUBÍ PUIG, A., «Responsabilidad civil e Inteligencia artificial. Un examen crítico de la propuesta de Directiva de 28 de septiembre de 2022», en M. Peguera Poch (coord.) *Perspectivas regulatorias de la inteligencia artificial en la Unión Europea*, Madrid, 2023, p. 255, son dos elementos los que han de darse para que el caso quede bajo el paraguas de la Propuesta de Directiva: por un lado, la negligencia como base de la responsabilidad; y por otro lado, la implicación de un sistema de IA en la causación del daño.

qué parte ha de soportar la carga de la prueba, qué grado de certeza se exige para que haya fuerza probatoria o cómo se define la culpa, *con excepción de lo previsto en los artículos 3 y 4*», que son precisamente los dos preceptos que introducen las medidas concretas para aligerar la carga de la prueba de la culpa y la causalidad⁸⁹⁸.

Por lo que se refiere a los casos de discriminación algorítmica, se podría plantear la duda de si tendrían que aplicarse las *presunciones refutables* previstas en los arts. 3 y 4 de la Directiva para aligerar al demandante la carga de probar la discriminación, o si debería aplicarse la *inversión de la carga de la prueba* recogida tanto en las Directivas comunitarias como en las normas nacionales de transposición. Considera un sector doctrinal al que nos adherimos, que la respuesta se encuentra en el art. 1.4 anteriormente transcrito. Esto es, si el enfoque de armonización mínima que adopta la PDRC permite que los demandantes puedan invocar normas más favorables del Derecho nacional que las previstas en la Directiva, no cabe duda de que en materia de no discriminación la norma más beneficiosa sería la que instaura la inversión de la carga de la prueba (recogida, entre otros preceptos, en el art. 30 de la Ley 15/2022 en relación con el art. 217.5 LEC), norma que además es perfectamente compatible con el Derecho de la Unión, tal como exige el precepto⁸⁹⁹. A mayor abundamiento, creemos que la aplicación preferente de la normativa de igualdad podría sustentarse en el *principio de especialidad*, en virtud del cual la ley especial prevalecería sobre la general en los supuestos contemplados en aquella norma.

1.3.3.3. Criterio de imputación de la responsabilidad y sujetos responsables

La PDRC se aparta del criterio de «gestión de riesgos» que sostenía la propuesta regulatoria del Parlamento y adopta la «culpa» como criterio de imputación de la responsabilidad. Por tanto, se centra en la persona que actuó de modo negligente como sujeto responsable, identificándose además la culpa con el propio incumplimiento material de una obligación exigible⁹⁰⁰.

⁸⁹⁸ En las restantes letras del art. 1.3 también se establece que la Directiva no afectará a: a) las normas del Derecho de la Unión que regulan las condiciones de responsabilidad en el ámbito del transporte; b) los derechos que puedan asistir a un perjudicado en virtud de las normas nacionales de transposición de la Directiva 85/374/CEE; c) las exenciones de responsabilidad y las obligaciones de diligencia debida establecidas en la Ley de servicios digitales.

⁸⁹⁹ Manteniendo esta opinión, NAVAS NAVARRO, S., «Régimen europeo en ciernes en materia de responsabilidad...», *cit.*, p. 35; EVANGELIO LLORCA, R., *op.cit.*, p. 599; o GIL MEMBRADO, C., «En el horizonte: la Directiva de responsabilidad civil extracontractual en materia de inteligencia artificial», en AA. VV., *Derecho y Medicina, desafíos tecnológicos y científicos*, (dir. C. GIL MEMBRADO), Madrid, 2023, p. 117.

⁹⁰⁰ Sobre este tema hace unas consideraciones interesantes, NAVAS NAVARRO, S., «Régimen europeo en ciernes en materia de responsabilidad...», *cit.*, p. 36. Según señala, el hecho de que en vir-

Se abandonan también los conceptos de *operador inicial y final* propuestos por el Parlamento y se acogen los términos de *proveedor* y *usuario* del sistema para mantener la coherencia con el RIA, al que remite para definirlos. No obstante, teniendo en cuenta que la versión definitiva del RIA ha sustituido el término «usuario» por el de «responsable del despliegue», las referencias que la PDRC hace a aquél deben entenderse hechas a este último.

Así, según el art. 3.3 RIA, el «proveedor» es la «persona física o jurídica o autoridad, órgano u organismo de otra índole públicos que *desarrolle* un sistema de IA o *un modelo de IA de uso general o para el que se desarrolle* un sistema de IA o *un modelo de IA de uso general y lo introduzca en el mercado o ponga en servicio el sistema de IA con su propio nombre o marca comercial*, previo pago o gratuitamente»⁹⁰¹. Por tanto, el proveedor podrá ser a su vez fabricante o no del sistema (pero las demandas se tendrán que dirigir contra el proveedor o personas que asuman las obligaciones del proveedor en base a esta Directiva). Por su parte, el «responsable del despliegue» se define en el art. 3.4 como «una persona física o jurídica o autoridad, órgano u organismo de otra índole públicos que *utilice un sistema de IA* bajo su propia autoridad, salvo cuando su uso se enmarque en una actividad personal de carácter no profesional».

De hecho, tal como dispone el art. 2.10 RIA respecto a su ámbito de aplicación, éste «*no se aplicará a las obligaciones de los responsables del despliegue que sean personas físicas que utilicen sistemas de IA en el ejercicio de una actividad puramente personal de carácter no profesional*». Ahora bien, tal como observa ATIENZA NAVARRO, esto no significa que los usuarios con fines privados no deban responder conforme al régimen de presunciones previsto, sino que sólo se les aplicará la presunción de la causalidad en circunstancias muy concretas. Particularmente, según el art. 4.6, cuando «haya interferido sustancialmente en las condiciones de funcionamiento del sistema de IA, o cuando tuviese la obligación y estuviese en condiciones de determinar las condiciones de funcionamiento del sistema de IA y no lo haya hecho»⁹⁰². *A sensu contrario*, cuan-

tud de la PDRC la prueba del comportamiento dañoso implique automáticamente dar por probado el criterio de imputación (la culpa), choca con los derechos nacionales de la responsabilidad civil o, como mínimo, con el derecho español. Se confunden así, según sigue diciendo, los presupuestos de la responsabilidad civil, ya que la diligencia (o la negligencia) pertenece al ámbito de la valoración jurídica y no de la infracción material. En su opinión, se trata de un cambio en las reglas de responsabilidad civil cuando intervienen sistemas de IA, que se basan en una concepción diferente de la «diligencia debida» en cuanto modelo de conducta, el cual implica el cumplimiento de una serie de deberes. Y en la p. 37 añade que, en materia de responsabilidad civil en el ámbito de la IA, la Propuesta implica un cambio en la concepción de la culpa en virtud del cual ésta dejará de valorarse jurídicamente.

⁹⁰¹ En un primer momento no se hacía referencia a «la gratuidad», lo que se añadió en la Orientación General del Consejo aprobada el 6 de diciembre de 2022.

⁹⁰² Dispone el Considerando 29 que este caso podría darse «cuando el usuario no profesional no respeta las instrucciones de uso u otras obligaciones de diligencia aplicables a la hora de elegir el

do no interfieran en el funcionamiento del sistema limitándose a usarlo de acuerdo a su finalidad, no se podrá presumir la causalidad.

También se emplea en diversas ocasiones a lo largo de la PDRC el término «operadores» que, según el art. 3.8 RIA, incluiría junto a proveedores y responsables del despliegue, al fabricante de productos, al representante autorizado del proveedor, al importador y al distribuidor.

Se cubre así el *ciclo de vida*⁹⁰³ del sistema por lo que a la responsabilidad se refiere, ya que, antes de su comercialización será el proveedor (o su representante autorizado⁹⁰⁴ o persona sujeta a las obligaciones del proveedor) el que responda del cumplimiento de los requisitos y obligaciones exigidas por el RIA para su puesta en circulación (art. 4.2 letras a-d), y una vez que el sistema se haya introducido en el mercado o puesto en servicio, los responsables podrían ser tanto el responsable del despliegue que incumpla las obligaciones recogidas en el art. 4.3, como el proveedor que obvie los deberes del art. 4.2 letra e)⁹⁰⁵.

No obstante, como precisa la doctrina, realmente la PDRC no hace referencia a «sujetos responsables» en tanto que no se determina cuál es el centro de imputación de la responsabilidad (como se hacía en la Propuesta del Parlamento de 2020 al señalar al «operador» como sujeto civilmente responsable)⁹⁰⁶. Por ende, habrá de entenderse que tanto los proveedores como los

ámbito de operación o de fijar las condiciones de funcionamiento del sistema de IA. Esto se entiende sin perjuicio del hecho de que el proveedor debe determinar la finalidad prevista de los sistemas de IA, incluidos el contexto y las condiciones específicas de uso, y eliminar o minimizar los riesgos de dicho sistema según convenga en el momento del diseño y el desarrollo, teniendo en cuenta los conocimientos y la experiencia de los usuarios previstos.

⁹⁰³ La versión dada al RIA por el Consejo en su Orientación General, añadía un art. 3.1 bis) donde se definía el «ciclo de vida» de un sistema IA como «la duración de un sistema de IA, desde su concepción hasta su retirada. Sin perjuicio de las competencias de las autoridades de vigilancia del mercado, dicha retirada puede producirse en cualquier momento de la fase de vigilancia poscomercialización por decisión del proveedor, e implica que el sistema no se podrá seguir utilizando. El ciclo de vida de un sistema de IA también finaliza si el proveedor u otra persona física o jurídica modifica sustancialmente el sistema de IA, en cuyo caso el sistema de IA sustancialmente modificado se considerará un nuevo sistema de IA». No obstante, el texto definitivo del RIA no ha incorporado este precepto.

⁹⁰⁴ Definido en el art. 3.5 RIA como «una persona física o jurídica ubicada o establecida en la Unión que haya recibido y aceptado el mandato por escrito de un proveedor de un sistema de IA o de un modelo de IA de uso general para cumplir las obligaciones y llevar a cabo los procedimientos establecidos en el presente Reglamento en representación de dicho proveedor».

⁹⁰⁵ Así lo mantiene, NAVAS NAVARRO, S., «Régimen europeo en ciernes en materia de responsabilidad...», *cit.*, p. 30.

⁹⁰⁶ En este sentido, entre otros, ATIENZA NAVARRO, M.L., «La responsabilidad civil por daños causados por inteligencia artificial...», *cit.*, p. 381; PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, p. 465; MARTÍ

responsables del despliegue de sistemas de alto riesgo serán los sujetos que principalmente puedan verse obligados o afectados por las medidas contempladas en la Propuesta para aligerar la carga de la prueba a las víctimas de los daños, y, aunque lo normal será que dichos sujetos sean los responsables de los daños, debe tenerse en cuenta que no serán los únicos posibles responsables porque cualquier persona física o jurídica podría ser demandada con base en las legislaciones nacionales sobre responsabilidad civil⁹⁰⁷.

Junto con estos sujetos, el art. 4.2 contempla también como posibles demandados a las «personas sujetas a las obligaciones del proveedor» remitiendo a los arts. 24 y 28.1 [de la Ley de IA] para su concreción. No obstante, tales preceptos fueron suprimidos y en la versión final del RIA hay que acudir al art. 25 para saber bajo qué condiciones una persona distinta al proveedor (concretamente, un distribuidor, importador, responsable del despliegue o tercero) puede ser considerada proveedor y quedar sujeta a las obligaciones de éste contempladas en el art. 16. Puesto que ya expusimos dichas condiciones en un momento anterior de esta obra, resumidamente recordaremos que se darán cuando los sujetos citados actúen como proveedores aparentes al poner su nombre o marca comercial en un sistema de IA de alto riesgo previamente introducido en el mercado o puesto en servicio, cuando modifiquen sustancialmente un sistema de IA de alto riesgo que ya haya sido introducido en el mercado o puesto en servicio, y cuando modifiquen sistemas (incluidos los de uso general) ya introducidos en el mercado o puestos en servicio que inicialmente no eran de alto riesgo pero que tras la modificación de su finalidad prevista pasan a serlo.

También deberán ser considerados proveedores y asumir sus obligaciones, los fabricantes de productos contemplados en la legislación de armonización de la Unión enumerados en la sección A del anexo I, a los que se incorpore un sistema de IA como componente, siempre que se cumpla alguna de las condiciones de dicho precepto.

GRAU, R., *op.cit.*, p. 9/14. Este último autor señala que la omisión en la PDRC de un listado de posibles sujetos responsables es coherente con la propia naturaleza de los sistemas de IA, en cuya cadena de producción hasta su definitivo funcionamiento pueden intervenir multitud de agentes. Sin embargo, para ATIENZA NAVARRO, la razón de esta omisión a la hora de fijar el centro de imputación de la responsabilidad es sencillamente que la PDRC no regula la responsabilidad.

⁹⁰⁷ ATIENZA NAVARRO, M.L., «La responsabilidad civil por daños causados por inteligencia artificial...», *cit.*, pp. 381-383. Afirma PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, p. 465, que si bien las reglas de la PDRC pueden aplicarse a cualquier sujeto al que el régimen interno de responsabilidad por culpa permita imputar el daño, su aplicación será residual cuando los demandados no sean proveedores, personas sujetas a las obligaciones del proveedor o usuarios de sistemas, ya que, respecto a la obligación de exhibir pruebas solo estarán obligados aquellos que dispongan de dicha información (que si bien no tienen por qué ser los sujetos mencionados, esto será lo más normal) y las normas sobre la presunción de la causalidad van dirigidas a proveedores (o asimilados) y usuarios.

Por último, en la medida en que la PDRC tampoco concreta cómo se debería distribuir la responsabilidad entre los causantes del daño cuando son varios, habrá que acudir al Derecho nacional para determinarlo. En el ordenamiento español, la responsabilidad en caso de pluralidad de deudores es, como regla general, mancomunada o parciaria. Por ende, la víctima debería dirigirse contra cada uno de los culpables para reclamar su parte de responsabilidad. No obstante, como es bien sabido, cuando no es posible determinar la parte en que cada sujeto contribuyó al daño (lo que en el ámbito de la IA no será un caso aislado ni excepcional), los tribunales suelen aplicar la solidaridad impropia o *in solidum* para proteger a los perjudicados. Sin perjuicio de dicha práctica, hubiera sido conveniente incluir una norma estableciendo la solidaridad de responsables, a fin de conseguir una mayor seguridad jurídica.

1.3.3.4. Medidas para aliviar la carga probatoria

La Comisión Europea parte de que las normas nacionales en materia de responsabilidad civil basadas en la culpa, no son adecuadas para tramitar las demandas por daños causados por sistemas de IA, ya que, con arreglo a dichas normas, las víctimas deben demostrar, en primer lugar, que ha habido una acción u omisión ilícita por parte de una persona que ha causado el daño, en segundo lugar, el daño propiamente sufrido, y por último, la relación de causalidad entre ambos. Sin embargo, las características específicas de la IA, incluidas su complejidad, autonomía y opacidad, pueden dificultar o hacer excesivamente costoso para las víctimas determinar quién es el responsable y probar que se cumplen tales presupuestos. En particular, puede resultar muy difícil demostrar que un dato de entrada concreto del que es responsable el demandado (o potencial demandado) ha dado lugar a una información de salida específica (o a que no se produjera dicha información de salida) que, a su vez, ha provocado el daño⁹⁰⁸.

Para facilitar esta carga de la prueba a las víctimas, la Propuesta introduce dos medidas principales: por un lado, un *derecho de acceso a la información necesaria* en poder de proveedores o responsables del despliegue de un sistema de alto riesgo para determinar qué ha provocado el daño y a qué persona física o jurídica se puede demandar por dicho daño (art.3); y por otro lado, *presumir iuris tantum la culpa y la causalidad en determinadas circunstancias* (arts. 3.5 y 4 respectivamente), lo que eximirá a las víctimas de tener que explicar detalladamente la manera en que determinada acción u omisión ha provocado el daño.

⁹⁰⁸ Considerando 3.

Estas medidas, cuyo objetivo es reparar el daño a la víctima y proteger su derecho a la tutela judicial efectiva⁹⁰⁹, servirán tanto para determinar el sujeto responsable, como para excluir a posibles demandados determinados erróneamente, ahorrando tiempo y costes a las partes y reduciendo la carga de trabajo de los tribunales.

Entre los diversos instrumentos disponibles para aliviar la carga de la prueba se opta por las presunciones *iuris tantum* por ser la herramienta menos gravosa e intervencionista para los Estados. Además, son frecuentes en los ordenamientos nacionales para ponderar los intereses de las partes en conflicto e incentivar el cumplimiento de las obligaciones de diligencia. Se descarta el establecimiento de otras medidas como la inversión de la carga de la prueba porque, según sostiene la Comisión, podrían frenar la innovación.

A) Primera medida: la exhibición de pruebas (art.3)

Mediante la obligación de exhibir las pruebas se trata de aliviar la carga probatoria del perjudicado de dos formas: por un lado, permitiendo que acceda a la información relevante para sustentar su demanda, y por otro lado, favoreciéndolo con el establecimiento de una presunción de culpa o negligencia del demandando en caso de que éste incumpla dicho deber de exhibición ordenado por el juez⁹¹⁰.

Se trata de una medida con la que se pretenden soslayar los problemas de opacidad algorítmica y mitigar la asimetría informativa entre las partes. Se aplica a proveedores, personas sujetas a las obligaciones de los proveedores y responsables del despliegue de sistemas de IA de *alto riesgo* pero no a los de no alto riesgo. El motivo reside en que, según el RIA, tales sistemas son los únicos que deben cumplir requisitos de documentación técnica, registro e información. Respecto a esta última circunstancia debe apuntarse que, aun cuando dichos sujetos no hubieran sido demandados, podrían quedar obligados a faci-

⁹⁰⁹ Tal como afirma la Comisión en la PDRC, «[s]i los retos de la IA dificultan demasiado el acceso a reparaciones, no existe un acceso efectivo a la justicia. Al garantizar una indemnización eficaz, estas normas contribuyen a la protección del derecho a la tutela judicial efectiva y a un juez imparcial, ambos reconocidos en la Carta de los Derechos Fundamentales de la Unión Europea»; y similar consideración hace en la PDPD al señalar que «[r]educir las restricciones a la hora de presentar reclamaciones y aligerar la carga de la prueba en casos complejos reforzarían el derecho a la tutela judicial efectiva, garantizado por el artículo 47 de la Carta de los Derechos Fundamentales de la Unión Europea».

⁹¹⁰ Como observa EVANGELIO LLORCA, R., *op.cit.*, p. 586, más que de una presunción en el sentido del art. 385 LEC, se trata de una sanción al demandado por incumplimiento de un deber impuesto por el juez. Aun cuando hace dicha consideración respecto a la presunción del art. 9.2 PDPD, entendemos que la misma es perfectamente extrapolable respecto a la presunción del art. 3.5 PDRC.

litar información a la víctima si disponen de ella⁹¹¹. En cambio, las solicitudes no pueden dirigirse a partes que no estén sujetas a obligaciones en virtud del RIA y que, por tanto, no tengan acceso a las pruebas.

Queda supeditada a dos condiciones: la primera, que el demandante o potencial demandante haya solicitado al proveedor o al responsable del despliegue que le permita acceder a la información que obra en su poder y éste se haya negado (arts. 3.1 y 3.2)⁹¹²; y la segunda, que la solicitud esté respaldada por hechos y pruebas suficientes para acreditar la viabilidad de la demanda. Si ambas condiciones se cumplen, los tribunales podrán ordenar la revelación de pruebas relacionadas con el daño generado por sistemas de IA de alto riesgo. No obstante, en caso de que el demandante no pudiera reunir esos hechos y pruebas para acreditar la viabilidad de su demanda o estos no se consideraran suficientes y, por ende, no pudiera ordenarse judicialmente la exhibición de pruebas a tenor de la PDRC, todavía podría conseguirse el acceso a esa información recurriendo a la distribución dinámica de la prueba respecto de los hechos sobre los que se considere que el demandado tiene mayor disponibilidad o facilidad probatoria *ex art.* art. 217.7 LEC, que seguiría siendo de aplicación con carácter general.

Si el demandado (o potencial demandado) incumple dicha orden, el órgano judicial presumirá su culpa, es decir, presumirá que ha incumplido alguno de los deberes de diligencia que le incumben y que dichas pruebas debían demostrar⁹¹³. No obstante, éste podrá destruir la presunción en su contra presentando pruebas que acrediten su diligencia.

⁹¹¹ En el Considerando 19 se explica, que esto es así porque «[p]odrían darse situaciones en las que las pruebas pertinentes para el asunto obren en poder de entidades que no sean parte en la demanda por daños y perjuicios, pero que estén obligadas a documentar o registrar dichas pruebas de conformidad con la [Ley de IA]. Por lo tanto, es necesario fijar las condiciones en que se puede ordenar a tales terceros con respecto a la demanda que exhiban las pruebas pertinentes». Según JORQUI AZOFRA, M., *Responsabilidad por los daños causados por productos y sistemas de inteligencia artificial*, Madrid, 202, p. 310, «[e]l demandante puede solicitar la exhibición de pruebas a proveedores o usuarios que no sean demandados sólo en el caso de que se hayan realizado sin éxito todos los intentos proporcionados para obtener las pruebas del demandado».

⁹¹² Respecto al demandante, el art. 3.2 dice literalmente que debe haber realizado «*todos los intentos proporcionados de obtener del demandado las pruebas pertinentes*». No bastará, parece, con solicitar sin éxito las pruebas a las personas mencionadas en el ap.1.

⁹¹³ Para RUBÍ PUIG, A., «Responsabilidad civil e Inteligencia artificial. Un examen crítico...», *cit.*, p. 274, la negativa a exhibir las pruebas solicitada por el juez estaría justificada cuando se demuestre razonablemente que los daños pueden imputarse objetivamente a una causa distinta del sistema de IA. No obstante, el art. 3.5 no precisa si la negativa del operador a dar información ha de ser justificada o no para aplicar la presunción de culpabilidad. Por tanto, entendemos que será indiferente este extremo a efectos de aplicar la presunción. Cuestión distinta sería que no procediera aplicar la presunción de causalidad si el demandado consigue probar con una probabilidad razona-

Esta disposición, recogida en el art. 3.5, tiene un doble objetivo: primero, promover la exhibición o el acceso a la información por parte del perjudicado (evitando así que entre en juego la presunción); y segundo, acelerar los procedimientos judiciales. Ahora bien, como observa atinadamente NAVAS NAVARRO, respecto al *potencial demandante* el derecho de acceso a las pruebas resulta poco práctico, ya que, si el mismo debe solicitarse judicialmente y para que se haga efectivo se deben aportar previamente hechos y pruebas suficientes para sustentar la viabilidad de la posible demanda, quizá le saldría más a cuenta interponer directamente una demanda en reclamación de daños (y tras ello, ya como demandante, pedir al juez que emita dicha orden) para evitar tener que demandar dos veces, una para solicitar al juez que ordene dicha exhibición y otra para pedir la compensación del daño⁹¹⁴.

Por lo que se refiere al aseguramiento de las pruebas obtenidas, el art. 3.3 establece que «[l]os Estados miembros velarán por que los órganos jurisdiccionales nacionales, a solicitud de un demandante, estén facultados para ordenar medidas específicas con el fin de conservar las pruebas mencionadas en el apartado 1». En nuestro Derecho, son los arts. 297-298 LEC los que regulan las medidas de aseguramiento de la prueba para evitar su destrucción o alteración.

Además, en aras de garantizar la proporcionalidad en la obligación de exhibición y conservación de las pruebas, la PDRC limita su alcance al mínimo necesario para demostrar el daño. Concretamente, la divulgación de la información probatoria y las medidas para su conservación, se ceñirán a «lo necesario y proporcionado» para sustentar la demanda (art 3.4), impidiendo solicitudes genéricas y pudiendo los tribunales tomar las medidas que consideren oportunas para garantizar la confidencialidad de los secretos comerciales durante y después del proceso judicial. Estas medidas deben incluir el acceso restringido a los documentos que contengan secretos comerciales y a las audiencias, documentos y sus transcripciones, a fin de equilibrar los intereses empresariales y de los afectados por la IA.

Precisamente este objetivo es el que, a juicio de la doctrina, se consigue con esta norma. Así, para MARTÍN CASALS, las condiciones bajo las que se propone esta medida permiten equilibrar la protección de las víctimas y el fomento de la innovación empresarial, ya que, si bien se eliminan en favor de las primeras los obstáculos para acceder a la información necesaria para probar sus alegatos, también se establecen ciertas garantías en favor de las segundas⁹¹⁵.

ble que su culpa no influyó en el resultado dañoso del sistema, ya que en este caso faltaría uno de los requisitos de aplicación (concretamente, el del art.4.1.b).

⁹¹⁴ NAVAS NAVARRO, S., «Régimen europeo en ciernes en materia de responsabilidad...», *cit.*, p. 33.

⁹¹⁵ MARTÍN-CASALS, M., «Las propuestas de la Unión Europea para regular la responsabilidad civil ...», *cit.*, p. 72.

A pesar de que, con carácter general, la utilidad de esta medida para aliviar la carga probatoria a las víctimas es considerable, se han puesto de relieve por la doctrina ciertas limitaciones de las que adolece en la práctica⁹¹⁶. La mayoría tienen que ver con el contenido mismo de la información objeto de exhibición, puesto que el art. 3.1 únicamente se refiere a «*pruebas pertinentes que obran en su poder*».

Como primera limitación se aduce, por un lado, el carácter inverificable del cumplimiento de esta obligación de los operadores, en tanto en cuanto a pesar de haber colaborado aportando información sobre el sistema, la misma podría ser poco útil o no todo lo pertinente o relevante que habría de ser para que el demandante pudiera probar su pretensión; y por otro lado, que hay cierta información que, aunque se quisiera transmitir al reclamante, hacerlo sería muy difícil por tratarse de información no codificable o cuya codificación y comunicación sería muy costosa.

La segunda limitación observada ha sido puesta de relieve en diversas ocasiones a lo largo de este trabajo y tiene que ver con la propia esencia del *aprendizaje automatizado* en que se basan los sistemas de IA. Teniendo en cuenta que las reglas generadas por el algoritmo van cambiando en función de los datos suministrados y de la probabilidad con la que ocurren los hechos en esa base de datos, será muy posible que la lógica decisional en el momento de producirse el daño y en el momento de exhibir la información sobre el mismo sea muy distinta. Por ello, se apunta que sería necesario concretar qué información habría que mostrar y si, en particular, habría de limitarse a la del momento de producción del «accidente».

En tercer lugar, se debería aclarar si sería suficiente con aportar una explicación general sobre la lógica aplicada por el sistema de IA, o si sería necesario informar sobre la lógica decisional concreta, es decir, sobre el concreto proceso decisional que ha conducido a un determinado resultado de salida⁹¹⁷. Lógicamente, la explicación de los motivos por los cuales el sistema acaba arrojando una determinada decisión sobre la persona, serán claves para poder valorar si interponer o no una demanda y en su caso articular bien la reclamación. A dicho cometido contribuye claramente el art. 86 RIA en virtud del cual toda persona afectada por una decisión basada en un sistema de IA de alto riesgo tiene derecho a *solicitar explicaciones claras y significativas sobre cómo el sistema de IA ha influido en el proceso de toma de decisiones y los principales elementos de dicha decisión*, con el fin de que las personas afectadas puedan ejercer sus

⁹¹⁶ Hasta seis retos del derecho a recibir información expone con detalle RUBÍ PUIG, A., «Responsabilidad civil e Inteligencia artificial. Un examen crítico...», *cit.*, pp. 279-287.

⁹¹⁷ Así lo estima GIL MEMBRADO, C., «Daños producidos por la IA ...», *cit.*, pp. 521-522.

derechos. Por tanto, la limitación apuntada habría quedado superada gracias a este nuevo derecho.

En cuarto lugar, se observa que en ocasiones no todos los sujetos que soportan la carga de exhibir información conocen el funcionamiento del sistema (por ejemplo, un responsable del despliegue no especializado). No obstante, como también explicamos en su momento, parece que respecto a los responsables del despliegue el art. 13 RIA se dirige precisamente a evitar dicha ignorancia y a garantizar que aquellos que utilicen sistemas de IA en su ámbito profesional estén informados adecuada y suficientemente sobre su funcionamiento, sobre los peligros que pueden implicar para los derechos fundamentales y sobre cómo interpretar las decisiones que arroje⁹¹⁸. De hecho, deben estarlo porque serán los que generalmente deban proporcionar dichas explicaciones a la persona afectada al ser los que generalmente adoptarán la decisión perjudicial⁹¹⁹.

Por otro lado, la multiplicidad de sujetos que quedan obligados por el deber de exhibir información plantea la cuestión de que, siendo varios a los que se reclame, sólo alguno cumpla la orden judicial de exhibir o conservar las pruebas que obran en su poder. Como dijimos, dicho incumplimiento permite presumir la culpa del demandado. Por tanto, se podría plantear si en estos casos dicha presunción decaería por el hecho de que uno solo de ellos hubiera colaborado, o si, por el contrario, bastaría la negativa de uno solo de ellos para que la presunción operara respecto de todos, o si sólo se aplicaría a quien no facilitó información debiendo hacerlo. A nuestro modo de ver, esta última opción sería la más lógica debiendo presumirse que, si son varios los que se niegan, concurriría culpa de todos ellos.

⁹¹⁸ Precisamente para asegurar que los responsables del despliegue conocen el sistema que van a implementar se impone que los sistemas de IA de alto riesgo deben ir acompañados de la documentación y las instrucciones de uso oportunas e incluir información clara y concisa, en particular sobre los posibles riesgos para los derechos fundamentales y riesgos de discriminación de las personas que puedan verse afectadas por el sistema en vista de su finalidad prevista, cuando corresponda. Y con el fin de conseguir que dichos operadores comprendan las instrucciones de uso, éstas deben contener ejemplos ilustrativos.

⁹¹⁹ No obstante, debe tenerse en cuenta aquí el pronunciamiento del TJUE en su sentencia de 7 de diciembre de 2023, asunto C 634/21 (OQ y Land Hessen, con intervención de SCHUFA Holding AG) –ya comentado en el apartado 2.5.2.1 del capítulo III–, que extiende a otros actores distintos del responsable del despliegue las obligaciones a cumplir con respecto a los interesados. En concreto, a los operadores que participen en la decisión automatizada pero sin adoptarla directamente (en el caso concreto, se trató de SCHUFA que era una entidad de información y análisis de riesgo). Extrapolando esto al ámbito de la IA, el proveedor quedaría obligado a proporcionar información sobre la decisión cuando la influencia del valor de probabilidad generado por él hubiera sido determinante en la decisión adoptada por el responsable de despliegue que afecta a la persona.

En quinto lugar, partiendo de la base de que el art. 3.1 se refiere a la exhibición de pruebas «*pertinentes*», habría que valorar qué información acerca del sistema podría cumplir esta característica. Para concretar este extremo, podría ser de utilidad la futura decisión que adopte el TJUE sobre la cuestión prejudicial pendiente de resolución en el ámbito de la protección de datos personales en el *asunto C-203/22 Dun & Bradstreet, planteada por el Verwaltungsgericht Wien (Austria) el 16 de marzo de 2022*, que gira sobre la utilización de perfiles para la evaluación de concesión de préstamos personales. En este caso se plantea al tribunal *qué debe entenderse por «información significativa» sobre el sistema en el sentido del art. 15.1 letra h) RGPD, así como si el carácter significativo de la información debe valorarse en el sentido de ser adecuada y suficiente para permitir al interesado ejercitar el resto de derechos que concede el RGPD, y cómo se ha de conjugar el derecho a la protección de datos con los derechos de terceros*⁹²⁰.

Aunque el caso aún no ha sido resuelto por el TJUE, el 12 de septiembre de 2024 el Abogado General De La Tour emitió unas conclusiones sobre el mismo. En su análisis sobre la interpretación del derecho del interesado a obtener «información significativa sobre la lógica aplicada» en procesos de toma de decisiones automatizadas, destacó que dicho derecho implica que se le proporcionen explicaciones suficientemente detalladas sobre el método utilizado para calcular una puntuación y las razones de un determinado resultado de forma que, con base en dicha información, pueda comprender el proceso que condujo a la decisión sobre su persona. Asimismo, la información significativa debe ser clara y accesible y, cuando sea necesario, debe completarse con explicaciones adicionales. Aunque se afirma también que dicha información no cubre el algoritmo utilizado, se considera que una descripción clara y comprensible de la lógica aplicada (el método utilizado, los criterios aplicados y su ponderación) es más beneficiosa para el interesado que el acceso mismo a la fórmula algorítmica⁹²¹. Además, según sigue diciendo, el interesado debe po-

⁹²⁰ El órgano jurisdiccional remitente pregunta al Tribunal de Justicia, en esencia, si, por una parte, el artículo 15, apartado 1, letra h), del RGPD debe interpretarse en el sentido de que la «información significativa sobre la lógica aplicada» en la adopción de una decisión automatizada incluye información suficientemente completa que permita al interesado verificar la exactitud de dicha información y su coherencia con la decisión de calificación de que se trata, incluido el algoritmo utilizado a efectos de dicha toma de decisiones automatizada. Por otra parte, si y, en su caso, en qué medida la protección de los derechos y libertades de otros, tales como la protección de los secretos comerciales invocada por el responsable del tratamiento, puede limitar el alcance del derecho de acceso del interesado contemplado en esta disposición.

⁹²¹ Vid., las Conclusiones del Abogado General SR. JEAN RICHARD DE LA TOUR presentadas el 12 de septiembre de 2024 disponibles en <https://curia.europa.eu/juris/document/document.jsf?text=&docid=290022&pageIndex=0&doclang=ES&mode=req&dir=&occ=first&part=1&cid=2084344>, donde afirma lo siguiente: «El responsable del tratamiento no está obligado a revelar al interesado información que, por su carácter técnico, es de una complejidad tal que no puede ser comprendida».

der verificar la exactitud de los datos utilizados y la decisión tomada. Es decir, debe poder evaluar si el tratamiento se basa en datos exactos y si el resultado o decisión se corresponde con tales datos⁹²². Apunta también que, aunque no es obligatorio *ex art. 15.1.h) RGPD*, el responsable podría ofrecer de forma anonimizada ejemplos de otras operaciones de tratamiento similares, a efectos de su comparación, para que el interesado pudiera comprender mejor la decisión automatizada de que ha sido objeto. Por último señala que, cuando la información que deba facilitarse al interesado en virtud del derecho de acceso del art. 15.1.h) RGPD pueda dar lugar a una vulneración de los derechos y libertades de otros (en particular por contener datos personales de terceros protegidos por el RGPD o bien un secreto comercial) dicha información debe comunicarse a la autoridad de control o al órgano jurisdiccional competentes para que, con pleno conocimiento de causa y respetando el principio de proporcionalidad, así como la confidencialidad de la referida información, puedan efectuar una ponderación entre los intereses en juego y determinar el alcance del derecho de acceso a la información que debe concederse al interesado.

En sexto y último lugar, se observa otro problema relacionado con la exhibición de pruebas que guarda conexión con el tiempo que proveedores o responsables del despliegue han de conservar la información pertinente. Esto es, si antes de que hayan podido determinarse los presupuestos de la pretensión indemnizatoria –principalmente el daño–, eliminaran los archivos de registro del sistema (teniendo en cuenta que sólo deben conservarlos el tiempo mínimo de 6 meses exigido por los arts. 19.1 y 26.6 RIA respectivamente para proveedores y responsables), podría ser imposible para la víctima probar dichos extremos.

Pues bien, junto con las limitaciones al derecho a la información o exhibición de pruebas observadas por la doctrina, a nuestro modo de ver la utilidad de este derecho también es cuestionable cuando se trata de casos de discriminación algorítmica, en la medida en que, presentados los indicios de discriminación se trasladaría al demandado la carga de la prueba y no sería necesario que el demandante presentara una solicitud de exhibición de prue-

por personas que no tengan conocimientos técnicos particulares, lo que permite excluir los algoritmos utilizados en la toma de decisiones automatizada» (ap. 96).

⁹²² En sus conclusiones también dispone, literalmente, lo siguiente: «esta información debe permitir al interesado ejercer los derechos que le garantiza el Reglamento 2016/679 y, en particular, el artículo 22. Por lo tanto, debe ser concisa, fácilmente accesible y fácil de entender, con un lenguaje claro y sencillo. Además, esta información debe ser suficientemente completa y estar suficientemente contextualizada para permitir al interesado verificar su exactitud y si existen una coherencia y una relación de causalidad objetivamente verificables entre, por una parte, el método y los criterios utilizados y, por otra, el resultado al que ha llegado la decisión automatizada en cuestión» (ap. 96).

bas al juez a fin de demostrar su culpa, ya que sería él quien debería probar su «no culpa», extremo que se traduciría en demostrar que la decisión no era discriminatoria o que aun siéndolo estaba justificada objetiva y razonablemente. Además, teniendo en cuenta que, según exige el art. 3.1 *in fine* de la PDRC, para solicitar al juez la exhibición de pruebas se requiere la aportación de «*hechos y pruebas suficientes que acrediten la viabilidad de la reclamación por daños*», además de que el proveedor, la persona sujeta a las obligaciones del proveedor o el responsable del despliegue se hubiera negado previamente a proporcionar información sobre el sistema a la víctima, cabe preguntarse si dicha exigencia vendría a coincidir con la que se exige en materia de discriminación para trasladar la carga de la prueba al demandado. Es decir, ¿se trataría de aportar ese principio de prueba o prueba *prima facie* de la vulneración del derecho fundamental? Si así fuera, el derecho a la exhibición de pruebas carecería de utilidad en los casos en que se alegue discriminación.

Teniendo en cuenta que, dada la complejidad y opacidad de los sistemas, esas sospechas fundadas de que existe discriminación tampoco serán fáciles de conseguir en el contexto algorítmico, la exhibición de pruebas sería útil en estos casos para conseguir reunir dichos indicios que permitan presumir que la decisión fue discriminatoria y no necesariamente para probar la negligencia del operador (objetivo al que se dirige este expediente). Quizá por ello, en los casos en que se alegue discriminación algorítmica como causa de los daños y partiendo de la complejidad de su prueba y de que sin una mínima cooperación del aparentemente responsable será extremadamente difícil reunir esos indicios, habría que rebajar el nivel de exigencia respecto de las condiciones para hacer efectivo este derecho de acceso a la información y entender: o bien que esos hechos y pruebas suficientes para respaldar la verosimilitud de la reclamación no han de ser de tal calado que puedan considerarse como verdaderos indicios para presumir que hay discriminación, sino de menor entidad (¿simples sospechas?)⁹²³; o bien que la mera negativa (extrajudicial) del operador a facilitar información sobre el sistema a la víctima (que debería, por su parte, realizar todos los intentos proporcionados de obtenerla *ex art.* 3.2) bastaría para pedir al juez que lo obligue a aportarla sin necesidad de acreditar la viabilidad de la demanda⁹²⁴. Recordemos que, según la doctrina sentada por el TJUE en el caso *Meister* ya comentado, «*la negativa de la empresa*

⁹²³ Sobre la diferencia entre *indicio* y *sospecha* y el estándar de prueba exigible para trasladar la carga de la prueba al demandado en los casos de discriminación tradicionales, me remito a lo explicado en el apartado 3.1.2 del capítulo III.

⁹²⁴ En caso de negarse a cumplir la orden judicial para exhibir o conservar las pruebas que obran en su poder, dicha negativa tendrá las consecuencias previstas en el art. 3.5. Concretamente, dicho órgano jurisdiccional nacional presumirá el incumplimiento por parte del demandado del deber de diligencia pertinente que las pruebas solicitadas estaban destinadas a probar.

a proporcionar dicha información a la solicitante, impidiéndole así reunir indicios suficientes para invertir la carga de la prueba, podía considerarse por los tribunales nacionales como uno de los factores que, junto con otros, permitieran presumir la existencia de discriminación».

Por último, habida cuenta de que, tal como ya se mencionó, la Directiva deja fuera de su ámbito de aplicación los supuestos en que se utilicen sistemas algorítmicos como soporte o apoyo a la toma de decisiones, parece que aquellos demandantes o potenciales demandantes que reclamen daños ocasionados por decisiones empresariales que se apoyen o tengan en cuenta el resultado algorítmico discriminatorio, no gozarían de este derecho de acceso a información relevante sobre el sistema. Siendo esto así, los perjudicados por decisiones discriminatorias semiautomatizadas no quedarían protegidos en base a la PDRC y no podrían utilizar este expediente para ser informados de cómo funcionan y deciden los sistemas. Ahora bien, teniendo en cuenta que el RIA se aplica tanto a decisiones plenamente automatizadas como a las semiautomatizadas (particular atención al derecho a una explicación del art. 86), y que el TJUE se ha posicionado recientemente acerca de ampliar las garantías del art. 22 RGPD también a las decisiones semiautomatizadas –exigencia que venía defendiéndose por la doctrina–, los perjudicados podrían obtener dicha información del sistema decisional y la concreta decisión que les afecta por otras vías.

Para garantizar la coherencia entre normativas, quizá debería ampliarse el ámbito de aplicación de este derecho de acceso a la información sobre el sistema también a los casos en que se usen como simple apoyo para la toma de decisiones finales en las que hay una «evaluación humana seguida de una acción u omisión humana» y el sistema se hubiera limitado a proporcionar información o asesoramiento que luego se tiene en consideración por el agente humano decisor. Más aun cuando la recomendación algorítmica tenga un peso determinante en la adopción de la decisión humana, ya que en tales casos la decisión podría considerarse como exclusiva o predominantemente automatizada según se desprende de la reciente doctrina del TJUE⁹²⁵.

⁹²⁵ Vid., sobre este asunto lo ya explicado al analizar la STJUE de 7 de diciembre de 2023. Crítica ARROYO AMAYUELAS, E., «El scoring de Schufa», *Indret*, n. 3, 2024, pp. 147, que se desprenda del razonamiento de esta sentencia que la decisión algorítmica (puntuación o valor de probabilidad) no estará cubierta por el art. 22 RGPD cuando no tenga una importancia decisiva en la decisión humana (en el caso del *scoring*, cuando los prestamistas atribuyan un peso significativo a otros factores para decidir la concesión o denegación del préstamo). En su opinión, esta postura no es acorde a la Directiva 2023/2225, de 18 de octubre de 2023, relativa a los contratos de crédito al consumo (DCC 2023), ya que su art. 18.8 ya no exige que la decisión del prestamista se base en un tratamiento exclusivamente automatizado, porque solo se refiere al «uso» de tratamientos automatizados de datos por su parte: «El precepto no distingue entre el modelo de negocio de las plataformas de préstamos

A nuestro juicio, el mero uso de sistemas de IA para la toma de decisiones con efectos en las personas, ya sean estos plena o semiautomatizados, debería implicar el nacimiento del derecho de información a los perjudicados (tal como ya reconoció hace unos años el legislador laboral en el art. 64.4 ET) porque ambos tipos de sistemas implican un importante riesgo de discriminación algorítmica por sesgos en el diseño o entrenamiento del algoritmo, entre otros factores. De hecho, por lo que se refiere a los sistemas semiautomatizados, puesto que su uso no está prohibido como regla general (al contrario de la prohibición que opera para los plenamente automatizadas *ex* art. 22, si bien con importantes excepciones), deben ser objeto de una mayor atención y regulación por los legisladores debido a su frecuencia de uso en prácticamente todos los sectores y actividades de la vida. Y es que, aunque determinar la relación de causalidad en dichos casos no sea más complicado para la víctima que hacerlo en situaciones en las que no interviene un sistema de IA (o eso afirma la Comisión)⁹²⁶, sólo será posible *detectar* la discriminación algorítmica en la toma de decisiones, sea ésta íntegra o semiautomatizada, si se puede tener acceso a la información sobre la decisión.

B) Segunda medida: la presunción de la relación de la causalidad (art. 4)

A diferencia del expediente relativo a la exhibición de pruebas, esta presunción es aplicable tanto a los sistemas de *alto riesgo* como a los de *no alto riesgo*, pero con importantes distinciones en uno u otro caso. Se trata de una presunción de causalidad entre la culpa del demandado y el comportamiento o decisión al que llega el sistema de IA (esto es, la información de salida o su

rápidos en línea en la que no hay intervención humana (e.g. *Buy Now, Pay Later*), esto es, propiamente, el modelo de decisiones automatizadas, y el modelo del prestamista tradicional que, antes de tomar decisiones personalizadas, se sirve de los inputs que le proporcionan las agencias de *scoring*. De acuerdo con la DCC 2023, basta con que la evaluación de la solvencia implique un procesamiento automatizado de datos personales, sin necesidad de que, además, la decisión deba considerarse adoptada automáticamente o se base de modo predominante en ese tratamiento automatizado». La DCC 2023 otorga relevancia a la valoración de la solvencia realizada por las agencias de *scoring* a petición del prestamista, independientemente del peso que tenga el *scoring* en la decisión final. Para esta autora, si el peso del *scoring* es relevante y ya tiene en sí mismo la naturaleza de decisión, «no debería ser posible que el carácter automatizado de la misma quedara desvirtuado por el hecho de que los procesos internos de la entidad bancaria que valoran la puntuación y adoptan una decisión final sobre el otorgamiento de crédito no sean a su vez automatizados. En tal caso, el interesado siempre debería tener derecho al suministro de información o a explicaciones adecuadas sobre la forma en que tal decisión se adoptó y sobre la lógica del algoritmo».

⁹²⁶ En estos casos parece claro que el responsable de la discriminación será el operador que adopta la decisión apoyándose en el resultado del sistema y no se podrá exonerar alegando el cumplimiento de las obligaciones del RIA, sino sólo probando que la decisión no se basó en motivos protegidos o que, aun siendo así, estaba justificada.

no producción que, a su vez, ocasiona el daño), con la que se quiere hacer frente al clásico problema de la causalidad jurídica o imputación objetiva que permite atribuir el daño a una persona⁹²⁷.

Debe señalarse, de entrada, que son tres condiciones («indicios o hechos base» en el ámbito de las presunciones) las que exige el art. 4.1 para la aplicación de la presunción *iuris tantum* que contempla: en primer lugar, que el demandante pruebe la culpa del demandado o de las personas de las que éste debe responder; o que el juez la haya presumido *ex art.* 3.5 (4.1.a); en segundo lugar, que pueda considerarse razonablemente probable atendiendo a las circunstancias del caso concreto que la culpa influyó en el resultado del sistema de IA (art. 4.1.b); en tercer lugar, que el demandante pruebe que el daño se debió al resultado del sistema (art.4.1.c).

Respecto a la primera condición, la existencia de culpa se equipara en la PDRC al incumplimiento de una serie de deberes de diligencia que para los sistemas de *alto riesgo* se recogen en los apartados segundo y tercero del art. 4 en función de si el demandado es el proveedor del sistema (o asimilado) o el responsable de su despliegue⁹²⁸.

En caso de que el demandado fuera el proveedor (o persona sujeta a las obligaciones del proveedor) se debería probar que el sistema de IA: a) No se ha desarrollado a partir de conjuntos de datos de entrenamiento, validación y prueba que cumplan los criterios de calidad del art. 10 RIA; b) No ha sido diseñado y desarrollado de modo que cumpla con los requisitos de transparencia del art. 13 RIA; c) No ha sido diseñado y desarrollado de modo que permita su vigilancia efectiva por personas físicas en el periodo en que se esté utilizando (art. 14 RIA); d) No ha sido diseñado y desarrollado de modo que a la luz de su finalidad prevista, alcance un nivel adecuado de precisión, solidez y ciberseguridad (art. 15 y art. 16 letra a RIA); e) No se han adoptado de forma inmediata las medidas correctoras necesarias para poner el sistema de IA en conformidad con los deberes de cuidado establecidos en el [capítulo III,

⁹²⁷ Así lo entienden, entre otros, PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, p. 475 o EVANGELIO LLORCA, R., *op.cit.*, p. 611. No obstante, coinciden en que este expediente no se aplicará a todos los casos en los que una IA cause daños y deban juzgarse con arreglo a un régimen subjetivo, sino principalmente a los daños causados por sistemas de alto riesgo y ni siquiera en todos ellos se aplicará de forma generalizada puesto que la Directiva reduce los supuestos en que se entiende que concurre culpa del demandado al incumplimiento de concretos deberes de diligencia que se enumeran en el texto.

⁹²⁸ El art. 2.9 PDRC define el «deber de diligencia» como la «norma de conducta exigida establecida por el Derecho nacional o de la Unión con el fin de evitar daños a los bienes jurídicos reconocidos a nivel nacional o de la Unión, incluidos la vida, la integridad física, la propiedad y la protección de los derechos fundamentales».

sección 2, del RIA] o para retirar el sistema del mercado o recuperar el sistema según proceda, de conformidad con el RIA (art. 16 letra j y art. 20).

Si el demandado fuera el responsable del despliegue del sistema de alto riesgo, el demandante deberá probar que: a) No cumplió con sus obligaciones de utilizar o supervisar el sistema de IA de conformidad con las instrucciones de uso adjuntas o, en su caso, de suspender o interrumpir su uso (art. 26 RIA); b) Expuso al sistema de IA a datos de entrada bajo su control que no eran pertinentes habida cuenta de su finalidad prevista (art. 26.4 RIA). No obstante, según se establece en el Considerando 26, la culpa de dicho sujeto podría determinarse no solo acreditando el incumplimiento de dichas exigencias específicas sino además acreditando el *incumplimiento de otras obligaciones de diligencia establecidas en el Derecho de la Unión o nacional* según establece el actual art. 26.3 RIA⁹²⁹, como por ejemplo, las que regulan el uso de la supervisión y la toma de decisiones automatizadas para el trabajo en plataformas o las que regulan el funcionamiento de aeronaves no tripuladas.

Además, teniendo en cuenta que se trata de una Directiva de mínimos, la doctrina entiende que nada impediría que los Estados pudieran añadir a este listado, otros deberes de diligencia cuyo incumplimiento implicase la culpa del demandando⁹³⁰. Por el momento, sobre la base de la versión final del RIA habría que añadir a dicho listado otras obligaciones recogidas en el art. 26, como la obligación de vigilar el funcionamiento del sistema, conservar los archivos de registro que los sistemas generen automáticamente en la medida en que dichos archivos estén bajo su control o cumplir las obligaciones de registro y la realización de evaluaciones de impacto sobre protección de datos en los casos que procedan⁹³¹.

⁹²⁹ Es su redacción finalmente aprobada dicho precepto dispone que: «Las obligaciones previstas en los apartados 1 y 2 no afectan a otras obligaciones que el Derecho nacional o de la Unión imponga a los responsables del despliegue ni a su libertad para organizar sus propios recursos y actividades con el fin de poner en práctica las medidas de supervisión humana que indique el proveedor».

⁹³⁰ Así, EVANGELIO LLORCA, R., *op.cit.*, p. 605.

⁹³¹ Se desprende del art. 26 RIA que las obligaciones de los responsables del despliegue son, en síntesis, las siguientes: 1) Adoptar medidas técnicas y organizativas necesarias para garantizar que los sistemas de IA se utilicen conforme a las instrucciones proporcionadas por el proveedor y en cumplimiento de la normativa aplicable; 2) Asignar la supervisión de los sistemas de IA a personas cualificadas que puedan intervenir y controlar su funcionamiento, asegurando el respeto de los derechos fundamentales y la seguridad de las personas; 3) Garantizar que los datos utilizados como entrada en los sistemas de IA sean pertinentes y suficientemente representativos para la finalidad prevista; 4) Monitorizar o supervisar el funcionamiento del sistema de IA y, si se detecta algún riesgo o incidente grave, informar sin demora al proveedor y a las autoridades competentes, y suspender el uso del sistema si es necesario; 5) Conservar los registros generados automáticamente por el sistema de IA durante un período adecuado a su finalidad, con un mínimo de seis meses, salvo que otra normativa disponga lo contrario; 6) Antes de poner en servicio o utilizar un sistema de IA de alto riesgo en el

Tales incumplimientos normativos se presumen negligentes *per se*. Por tanto, probados por el demandante (o no contrarestados por el demandado en los supuestos de discriminación), se entenderá acreditada su culpa.

En caso de que el demandante no consiguiera probar la negligencia del demandado (o en los casos de discriminación, que el demandado consiguiera probar el cumplimiento de los requisitos previstos en las secciones 2 y 3 del Capítulo III del RIA u otras obligaciones de diligencia fijadas por el Derecho nacional o comunitario), quedaría excluida su responsabilidad en tanto en cuanto en los sistemas de responsabilidad subjetiva o por culpa es lógico que la ausencia de culpa excluya la responsabilidad. No obstante, como apunta NAVAS NAVARRO, el cumplimiento de dichos requisitos «no tiene porqué significar necesariamente que la conducta del demandado fuera diligente, si el modelo de conducta requería extremar los cuidados más allá de lo establecido en esos capítulos. Pero esto no podrán –parece– a tenor de la Propuesta valorarlo los tribunales»⁹³². A lo que podría añadirse, en nuestra opinión, que la prueba del cumplimiento de dichos requisitos tampoco garantizaría siempre y en todo caso que el sistema no arrojará decisiones discriminatorias (sino simplemente que los operadores no tendrían la culpa de ello porque cumplieron las obligaciones debidas). Cabría preguntarse, por tanto, quien debería responder (si es que alguien debería hacerlo) en caso de que, incluso respetando todos los requisitos legales y obligaciones impuestas a proveedores, responsables del despliegue y otras partes, el sistema generase daños (entre otros muchos casos, por llegar a resultados sesgados con efectos discriminatorios)⁹³³.

Respecto a la segunda de las condiciones del art. 4.1, se infiere del apartado b) que no toda culpa debe dar lugar a la aplicación de esta presunción, sino que la misma solo procederá cuando pase el test de lo «razonablemente probable». Se entiende la culpa como el incumplimiento de un deber de di-

lugar de trabajo, informar a los representantes de los trabajadores y a los empleados afectados sobre su uso; 7) Si el responsable del despliegue es una autoridad pública o una institución de la Unión Europea, cumplir con las obligaciones de registro del sistema establecidas en el art. 49 RIA; 8) Utilizar la información proporcionada por el proveedor para realizar una evaluación de impacto relativa a la protección de datos, según lo establecido en el RGPD; 9) Cooperar con las autoridades competentes, proporcionando la información y documentación necesaria para demostrar la conformidad del sistema con el RIA; 10) En el caso de que se empleen sistemas de IA de alto riesgo del anexo III que *tomen decisiones o ayuden a tomar decisiones* relacionadas con personas físicas, *deberán informar a dichas personas de que están expuestas a la utilización de tales sistemas*, sin perjuicio de lo dispuesto en el art. 50 sobre obligaciones de transparencia específicas cuando se usen determinados sistemas IA destinados a interactuar con personas físicas. A todas estas obligaciones se suma el cumplimiento de unos estrictos requisitos cuando se utilicen sistemas de identificación biométrica remota en diferido en el marco de investigaciones penales (vid., art. 26.10).

⁹³² «Régimen europeo en ciernes...», *cit.*, p. 37.

⁹³³ HUERGO LORA plantea si deberían ser las autoridades nacionales de supervisión.

ligencia *destinado directamente a proteger contra los daños sufridos*, lo que no implica más que la exigencia de que concurra el criterio de imputación objetiva del «fin de protección de la norma de cuidado». Es por ello que no será suficiente con que el sujeto responsable haya actuado de forma negligente sino que, además, esa negligencia deberá ser relevante en relación con el daño sufrido por la víctima⁹³⁴. La propia Comisión ejemplifica esta cuestión en el Considerando 25 al señalar que, «puede considerarse razonablemente probable que la culpa ha influido en la información de salida o en la no producción de una información de salida cuando dicha culpa consista en el incumplimiento de un deber de diligencia con respecto a la limitación del perímetro de funcionamiento del sistema de IA y los daños se hayan producido fuera del perímetro de funcionamiento. Por el contrario, no puede considerarse razonablemente probable que el incumplimiento de la obligación de presentar determinados documentos o de registrarse ante una autoridad determinada –aunque ello pueda estar previsto para esa actividad concreta o incluso ser expresamente aplicable al funcionamiento de un sistema de IA– ha influido en la información de salida producida por el sistema de IA o en la no producción de dicha información de salida por sistema de IA».

En concreto, la relevancia de la culpa en la producción del daño deberá alcanzar el estándar de lo «razonablemente probable», lo que, para un sector doctrinal supone *rebajar el estándar de prueba* de la influencia de la culpa del demandado en la causación del daño⁹³⁵. De esta forma, no será preciso para

⁹³⁴ Así, EVANGELIO LLORCA, R., *op.cit.*, pp. 606-607 y PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, p. 472. Como señala también ATIENZA NAVARRO, M.L., «Distintos modelos para la responsabilidad civil por daños causados por la Inteligencia Artificial. A modo de comparación entre distintas propuestas de la Unión Europea», en M. Peguera Poch (coord.), *Perspectivas regulatorias de la inteligencia artificial en la Unión Europea*, Madrid, 2023, p. 334, en ambos supuestos (prueba o presunción de culpa ex art.3.5), la culpa debe estar conectada causalmente con el daño producido, de forma que si la negligencia consiste en no aportar una serie de documentos relacionados con una autorización administrativa requerida para poner en funcionamiento el sistema y el daño causado a la víctima ha sido un daño físico que nada tiene que ver con dicha autorización, no podrá presumirse la relación de causalidad; y explica PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, p. 469, que «esta relación entre la culpa y el daño se ha expresado tradicionalmente mediante el binomio previsibilidad-evitabilidad desde el punto de vista del estándar de diligencia. Así para que haya culpa, además de la presencia de una conducta incorrecta (contraria al estándar de diligencia), es necesario que el daño sea previsible para una persona diligente y que, además, sea susceptible de ser evitado ejecutando la conducta correcta».

⁹³⁵ Así, EVANGELIO LLORCA, R., *op.cit.*, p. 606 y PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, p. 472. Para este último, «lo que propone el art. 4.1 de la PDRIA es rebajar el estándar probatorio aplicable a la prueba de la previsibilidad/evitabilidad del daño, sustituyendo el ordinario (cualquiera que éste sea en el Derecho interno), por otro atenuado». En sentido contrario, MARTÍN-CASALS, M., «Las propuestas de la Unión Europea para

entender probado el nexo causal entre la culpa y el daño que el juez esté convencido de ello «más allá de toda duda razonable» (estándar de prueba que se exige en la mayoría de ordenamientos de Derecho continental) sino que será suficiente con que considere que hay una *probabilidad razonable* de que sea así. No obstante, teniendo en cuenta que en virtud de su art. 3.1 la Directiva no pretende armonizar las normas nacionales sobre la carga y estándar de la prueba, será cada Estado miembro el que habrá de determinar qué grado de certeza se exigirá –entre el comportamiento negligente del demandado y el resultado dañoso del sistema– para que haya fuerza probatoria⁹³⁶.

Sobre si debe ser el demandante el que haya de acreditar esa «probabilidad razonable» a la que se refiere el art. 4.1. b), la doctrina se posiciona por una respuesta negativa sobre la base del tenor literal de dicho apartado y su comparación con el de las otras dos letras del art. 4.1 (a y c), ya que en estas últimas sí se atribuye expresamente al demandante la carga de la prueba que se omite en la primera. Además, si fuera así, al demandante no le aprovecharía la presunción porque en realidad se le estaría exigiendo la prueba del nexo causal⁹³⁷. Parece, por tanto, que dicha situación habría de ser evaluada por el juez atendiendo a las circunstancias del caso. No obstante, aun así, esta valoración también restaría utilidad a la propia presunción⁹³⁸.

regular la responsabilidad civil ...», *cit.*, p. 73, considera que «no se trata aquí de permitir al juez utilizar un estándar de prueba reducido ('razonablemente probable', parecido al *more probable than not* angloamericano) –que en nuestra doctrina a menudo se confunde con la «causalidad probabilística» (*proportional liability*), con la que nada tiene que ver– sino de permitirle presumir el nexo causal entre la actuación culposa y la información o falta de información del sistema de IA en atención a las circunstancias del caso».

⁹³⁶ Considera PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, pp. 472-473, que esa variación en la reducción del estándar de prueba dependerá, por un lado, del estándar de partida que se exija en cada ordenamiento, y por otro lado, de la generosidad con la que cada jurisdicción nacional interprete lo que debe entenderse por «probabilidad razonable». Así, respecto al primer extremo explica que «Si, por ejemplo, en el ordenamiento interno se aplica el estándar de la preponderancia, de corte anglosajón –para entender probado un hecho, su existencia debe ser más probable que si inexistencia– (una probabilidad >50%), entonces la norma de la PDRIA deberá conducir a aplicar probabilidades por debajo del 50%. Mientras que si el estándar de partida es la certeza del hecho (digamos un 80% de probabilidad), la PDRIA simplemente obligará a rebajar esa elevada probabilidad». Como conclusión sobre la utilidad de la medida contemplada en la Propuesta de Directiva opina dicho autor, que «En el fondo, todo se reduce a reducir en un porcentaje poco claro la intensidad con la que debe probarse que la negligencia del demandado tuvo que ver con la actuación dañosa de la IA. Al final, serán realmente los órganos jurisdiccionales los que determinen, con su interpretación de lo que signifique “razonablemente probable” los que podrán marcar, o no, la diferencia para las víctimas».

⁹³⁷ NAVAS NAVARRO, S., «Régimen europeo en ciernes en materia de responsabilidad...», *cit.*, p. 38.

⁹³⁸ Se pregunta NAVAS NAVARRO, S., «Régimen europeo en ciernes en materia de responsabilidad...», *cit.*, p. 38, de qué sirve la presunción *iuris tantum* si el tribunal puede decidir no aplicarla porque considera que no existe –o no existe razonablemente– el nexo causal.

Por estas razones, la doctrina concluye que hubiera resultado más útil para el perjudicado que se exigiera solo la prueba de la culpa y del nexo causal entre el resultado del sistema y el daño, sin necesidad de valorar si ese nexo de causalidad que se pretende presumir es razonablemente probable que existiera. Según observa NAVAS NAVARRO, «da la impresión de que el legislador comunitario no estaba muy convencido de la equiparación legal entre culpa y comportamiento omisivo del demandado, pues si, a pesar de éste, no se puede considerar “razonablemente probable” que haya influido en el resultado –o en su ausencia– producido por el sistema de IA, no podrá aplicarse la presunción *iuris tantum* del nexo causal. Así, lo que entró por la puerta se acaba echando por la ventana»⁹³⁹.

En tercer y último lugar, el art. 4.1.c) exige que el demandante demuestre que fue la información de salida producida por el sistema de IA o su no producción, lo que causó los daños. Es decir, debe probarse el nexo causal entre el resultado del sistema (o su falta) y el daño ocasionado. En los casos objeto de nuestro estudio, esto implicaría probar que la decisión del sistema fue lo que generó daños a la persona o grupo de personas al vulnerar su derecho a la igualdad y no discriminación.

Por otro lado, se establece una excepción a la aplicación de la presunción de causalidad en el art. 4.4 cuando, en relación con un sistema de *alto riesgo*, el demandado demuestre que el demandante pudo acceder razonablemente a pruebas y conocimientos especializados suficientes para demostrar el nexo causal. Esto podría ocurrir, por ejemplo, cuando el demandado hubiera exhibido al demandante las pruebas que se le solicitaron extrajudicialmente, o cuando el demandante fuera el responsable del despliegue que reclama una indemnización al proveedor, ya que, en estos casos, debido a las obligaciones de documentación, registro, transparencia y comunicación de información que el proveedor debe cumplir respecto a él, éste contaría ya con la información del sistema. En tales casos, cuya concurrencia deberá probar el demandado, los tribunales no deberían aplicar la presunción debiendo probarse el nexo causal por el demandante.

Esta excepción sería un incentivo para que los demandados cumplieran sus obligaciones de exhibición de pruebas y las medidas del RIA dirigidas a garantizar un alto nivel de transparencia y los requisitos de documentación y registro de los sistemas.

Por otro lado, por lo que se refiere a los *sistemas de IA que no son de alto riesgo* y con independencia de que haya sido demandado el proveedor (o asimila-

⁹³⁹ Concluye esta autora afirmando que la PDRC alivia poco la carga de la prueba y lo hace forzando el régimen de la responsabilidad por culpa para evitar acudir al régimen de responsabilidad objetiva.

do) o el responsable del despliegue, el art. 4.5 establece que la aplicación de la presunción de causalidad sólo se aplicará cuando el órgano jurisdiccional nacional considere *excesivamente difícil* para el demandante demostrar el nexo causal entre la culpa y los resultados del sistema⁹⁴⁰. Esa dificultad deberá evaluarse a la luz de las características de determinados sistemas de IA, como la autonomía y la opacidad, características que al no ser exclusivas de los sistemas de alto riesgo, hacen que sea criticable la diferencia de trato que la Comisión establece para aplicar la presunción de la causalidad en unos casos y otros⁹⁴¹.

Como regla general, para los supuestos de daños causados por sistemas que no sean de alto riesgo se aplicarán en principio las reglas ordinarias de la carga de la prueba, debiendo probar el demandante el nexo causal entre la culpa y el daño (culpa que aquí se ceñiría al incumplimiento de las obligaciones de transparencia –art.50 RIA– o los códigos de conducta voluntarios existentes –art. 95 RIA–), sin perjuicio de que también tendría que probar el daño y el nexo causal entre éste y el resultado del sistema⁹⁴². Esta regla general se vería excepcionada en los casos en que la opacidad y autonomía del sistema complicaran de forma «excesiva» al demandante la prueba del nexo causal, lo que a nuestro juicio sucederá con bastante frecuencia.

Por último, cuando las demandas se dirijan contra usuarios no profesionales, la presunción de causalidad sólo podrá aplicarse si, además de acreditarse los hechos del apartado 1, se dieran las circunstancias previstas en el art. 4.6, esto es, cuando el demandado hubiera interferido sustancialmente en el funcionamiento del sistema de IA o cuando tuviese la obligación y estuviese en condiciones de determinar las condiciones de funcionamiento y no lo hubiera hecho. Esta diferencia de trato respecto a los usuarios profesionales se explica en el Considerando 29, según el cual, el aligeramiento de la carga de probar la causalidad no siempre es apropiado cuando el demandado es una persona que utiliza el sistema de IA para sus actividades privadas: «En tales circunstancias, a fin de lograr un equilibrio entre los intereses de los perjudicados y los de los usuarios no profesionales, *ha de tenerse en cuenta si dichos usua-*

⁹⁴⁰ No obstante, apunta NAVAS NAVARRO, S., «Régimen europeo en ciernes en materia de responsabilidad...», *cit.*, p. 40, que para aplicar la presunción, además de ser excesivamente difícil probar el nexo causal, también tendría que considerarse *razonablemente probable* que la culpa ha influido en los resultados producidos.

⁹⁴¹ Para PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, p. 475, tan complicada podrá ser su prueba en uno u otro caso puesto que «lo que hace que la prueba sea difícil no es el nivel o la intensidad del riesgo creado por la IA de turno, sino el modo autónomo, opaco e impredecible con el que opera esta tecnología».

⁹⁴² En mismo sentido, NAVAS NAVARRO, S., «Régimen europeo en ciernes en materia de responsabilidad...», *cit.*, p. 40 y EVANGELIO LLORCA, R., *op.cit.*, p. 609.

rios no profesionales pueden aumentar, mediante su comportamiento, el riesgo de que un sistema de IA cause daños». En base a esto, podrían distinguirse varios supuestos:

- a) Aquéllos en que los usuarios no profesionales adquieran sistemas de IA y los utilicen conforme a su finalidad y sin interferir sustancialmente en su funcionamiento. A estos casos no se les aplicará la presunción de causalidad, ya que, si el usuario se ha limitado a adquirir un sistema de IA –comercializado por considerarse suficientemente seguro al cumplir las exigencias legales correspondientes– y a ponerlo en funcionamiento de forma adecuada, no estaría incrementando el riesgo de causar daños inherente al sistema.
- b) Aquéllos en que los usuarios interfieran sustancialmente en las condiciones de funcionamiento del sistema de IA o cuando el usuario, teniendo la obligación de determinar las condiciones de funcionamiento del sistema de IA y estando en condiciones de hacerlo, no lo hubiera hecho. Aquí se aplicará la presunción de causalidad, siempre que se cumplan todas las demás condiciones. Este caso podría darse cuando el usuario no profesional no hubiera respetado las instrucciones de uso u otras obligaciones de diligencia aplicables a la hora de elegir el ámbito de operación o de fijar las condiciones de funcionamiento del sistema de IA.

En definitiva, como apunta entre otros RUBÍ PUIG, solo se aplicará la presunción cuando, por ejemplo, el usuario no profesional hubiera incrementado el riesgo del sistema al manipular su funcionamiento o alterar sustancialmente las condiciones de uso, o hubiera infringido una posición de garante definida legal o contractualmente en relación con el funcionamiento del sistema⁹⁴³. Esto se entiende sin perjuicio de que es el proveedor quien debe determinar la finalidad prevista de los sistemas de IA (incluidos el contexto y las condiciones específicas de uso) y eliminar o minimizar sus riesgos teniendo en cuenta los conocimientos y experiencia de los usuarios que previsiblemente utilizarán el sistema.

Finaliza el art. 4 PDRC con una declaración expresa del carácter *iuris tantum* de la presunción que contiene, disponiendo su apartado séptimo que «Al demandado le asistirá el *derecho de refutar* la presunción establecida en el apartado 1». Por tanto, podrá desvirtuarla si consigue probar que el daño no derivó del incumplimiento de su concreto deber de diligencia, sino de una causa distinta (por ejemplo, un caso fortuito, la intervención de un tercero, etc). En opinión de PEÑA LÓPEZ, sería conveniente exigir al demandando

⁹⁴³ RUBÍ PUIG, A., «Responsabilidad civil e Inteligencia artificial. Un examen crítico...», *cit.*, p. 294.

lo mismo que al demandante, es decir, «que existe otro motivo o causa que permita explicar, con una probabilidad razonable, la actuación de la IA. Esta interpretación no sólo situaría a las partes en una posición similar ante las peculiaridades de esta tecnología, sino que, al final, permitiría al juez decidir el asunto ponderando cuál de las dos explicaciones del daño (la negligencia del demandado o la causa alternativa alegada por éste) le parece que es “razonablemente” la más probable»⁹⁴⁴.

Por último, debe quedar claro que las presunciones previstas en la PDRC no eximirán al demandante de tener que probar la existencia y alcance de los daños que se reclamen. A estas cuestiones dedicaremos el siguiente apartado.

2. LOS DAÑOS EN EL ESCENARIO DE LA IA Y LOS MECANISMOS PARA SU REPARACIÓN

2.1. EL DAÑO COMO ELEMENTO INDISPENSABLE DE LA RESPONSABILIDAD CIVIL: ¿NOVEDADES RESPECTO A LOS DAÑOS DERIVADOS DEL USO DE LA IA?

El daño es un requisito esencial de toda acción de reclamación de responsabilidad civil, por tanto, si no hay daño no hay obligación de responder. Además, como normal general, el daño no se presume y la carga de su prueba corresponde al demandante conforme al art. 217.2 LEC.

Los supuestos de daños ocasionados por sistemas inteligentes no suponen una excepción a lo afirmado. Por tanto, la existencia y prueba del daño también se erige en presupuesto básico de la pretensión indemnizatoria⁹⁴⁵. No

⁹⁴⁴ PEÑA LÓPEZ, F., «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil ...», *cit.*, p. 474. En su opinión, «la eficacia de la presunción no sólo depende de lo que el juzgador considere una probabilidad razonable de influencia de la culpa, sino también del rigor con el que se interprete el requisito de la prueba en contrario». Mantiene que si se interpreta en el sentido de que, para desvirtuar la presunción del art. 4.1, basta con que el demandado pruebe que el comportamiento del sistema de IA era imprevisible e inevitable para él (lo que podrían demostrar sin excesiva dificultad teniendo en cuenta precisamente que estos rasgos son inherentes a muchos sistemas de IA –sobre todo a los de *deep learning*–), la posición de las víctimas quedaría debilitada y la presunción sería inoperante.

⁹⁴⁵ Como ha afirmado el TS en sentencia 141/2021, 15 de marzo 2021 (RJ 807/2021), «[e]l daño es el presupuesto para la existencia de la responsabilidad civil, en tanto en cuanto marca sus límites y el *quantum indemnizatorio*». De hecho, el daño cumple tres funciones básicas en el sistema de responsabilidad civil extracontractual: pone en marcha el propio sistema de reparación y compensación, determina de qué debe responder su causante y precisa en cuánto debe valorarse su responsabilidad.

obstante, el hecho de que un determinado comportamiento o actividad cause daños, no significa que siempre y en todo caso deban ser resarcidos, sino que sólo lo serán cuando sean *jurídicamente relevantes* en el sentido de ser ciertos y efectivos, no futuros e inciertos⁹⁴⁶. Además, el daño resarcible ha de ser económicamente evaluable e individualizado, pudiendo constituir una lesión física o mental o bien un daño material o inmaterial (por ejemplo, pérdidas financieras o económicas, pérdida de empleo o de oportunidades educativas o sanitarias, restricción indebida de la libertad de elección o de expresión, pérdida de la privacidad y cualquier infracción del régimen jurídico aplicable que sea perjudicial para una persona)⁹⁴⁷.

Estas ideas han de tenerse muy presentes en el tema que nos ocupa, ya que los daños ocasionados por la IA en general, y por la discriminación algorítmica en particular⁹⁴⁸, presentan una problemática concreta por varios motivos: primero, porque *no suelen ser suficientemente individualizables*, esto es, no siempre afectan a personas concretas de manera clara y directa, sino que suelen tener un efecto sistémico afectando a grupos de personas o instituciones en su conjunto, lo que dificulta la identificación de víctimas individuales o concretas; segundo, porque aunque puedan individualizarse, suelen ser *mínimos o ni siquiera resarcibles*. Por ejemplo, el daño que pueda llegar a sufrir una persona al comprobar que una determinada red social le muestra anuncios para determinados puestos de trabajo y no para otros en función exclusivamente de su género o raza, puede no conllevar mayores consecuencias que el disgusto del usuario concreto u otras emociones como ira o frustración que en sí mismas no son indemnizables al formar parte de los riesgos generales de la vida⁹⁴⁹. Además, el impacto

Así lo explica, GÓMEZ LIGÜERRE, C. «Concepto de daño moral», en GÓMEZ POMAR, F. y MARÍN GARCÍA, I. (dirs.), *El daño moral y su cuantificación*, Madrid, 2023, p.33.

⁹⁴⁶ YZQUIERDO TOLSADA, M., *Responsabilidad civil extracontractual. Parte general*, Madrid, 2020, p. 195, en base a la delimitación de García Serrano, afirma que ha de separarse el «concepto de daño» y el «daño en sentido jurídico». Asimismo, dentro de este debe distinguirse entre el daño a bienes patrimoniales y a bienes no patrimoniales, admitiendo a su vez dentro de este último, los daños físicos a la persona y los daños morales propios.

⁹⁴⁷ TAPIA HERMIDA, A., «La responsabilidad civil derivada del uso de la inteligencia artificial y su aseguramiento», *Revista de la Asociación Española de Abogados Especializados en Responsabilidad Civil y Seguro*, n. 76, 2020, p. 127. GÓMEZ LIGÜERRE, C. «Concepto de daño moral», cit., p. 35, indica que el daño es cualquier pérdida de utilidad que sufre la persona en sus bienes materiales o inmateriales y siempre que resulte cierto y sea valorable económicamente debe indemnizarse cuando concurren, además, el resto de presupuestos que exigen las normas de responsabilidad civil extracontractual.

⁹⁴⁸ Respecto a la discriminación algorítmica, un sector doctrinal considera que habría que hablar de «riesgos», más que de «daños». En este sentido, explica RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», cit., p. 631, que mientras el daño es un hecho cierto, el riesgo es un impacto negativo incierto pero cuya probabilidad se conoce.

⁹⁴⁹ No obstante, siendo esto cierto, también lo es que en otros casos, no infrecuentes, los daños ocasionados a la víctima de la discriminación algorítmica pueden ser considerables y traducirse

individual es difícil de medir en comparación con el efecto generalizado que pueda tener sobre muchas personas debido a la escala y velocidad con la que dichas «microdiscriminaciones» pueden reproducirse y acabar amplificándose y perpetuando estereotipos; por último, porque sus *consecuencias jurídicas no suelen ser inmediatas, sino que se van manifestando con el tiempo*, lo que implica que el daño puede tardar en ser reconocido por las víctimas y, por ende, en ser resarcido.

Por éstas y otras razones la doctrina considera que muchas de las externalidades negativas asociadas a los sistemas de IA no son el tipo de menoscabos que el Derecho de daños persigue prevenir y compensar, es decir, no son daños indemnizables⁹⁵⁰. De hecho, al Derecho de daños no le interesan todos los riesgos y/o daños que puedan derivarse de una actividad o sector, como puede ser la IA, y muchos de ellos se tratan de encauzar *ex ante* a través de normas de seguridad que tratan de prevenirlos y minimizarlos.

Ahora bien, si a pesar de tales medidas de seguridad los daños acaecen, se hace preciso determinar si son indemnizables o no, cuestión que a falta de normas armonizadas debe solucionarse acudiendo a los regímenes nacionales. En este sentido, mientras que en algunos ordenamientos no se indemnizan determinadas categorías de daños que pueden ser frecuentes en el ámbito de la IA (por ejemplo, los morales puros), ni daños de escasa cuantía (mínimos o de bagatela) y suelen ponerse límites a los perjuicios que se presentan de forma dispersa y alcanzan a un grupo más o menos indeterminado de víctimas, en España contamos con un sistema abierto de reparación del daño en virtud del cual cualesquiera daños son en principio resarcibles si concurren los presupuestos para ello⁹⁵¹. Por ende, desplazada la regla general informada por el principio *casum sentit dominus*, la norma será la *reparación integral del daño*⁹⁵².

Siendo esto así, en nuestro ordenamiento jurídico encontraremos menos problemas para resarcir los daños ocasionados por la IA que en otros sistemas donde rige el principio de tipicidad y sólo unas categorías de daños son indemnizables con arreglo a las normas de la responsabilidad civil extracontractual. Tampoco será especialmente relevante que los daños sean consecuencia

en pérdidas de oportunidades con mucha repercusión y trascendencia en la vida de la persona (piénsese en la negativa a acceder a un centro educativo prestigioso, a un crédito para la compra o alquiler de una vivienda, etc).

⁹⁵⁰ Los motivos para mantener esta afirmación los desarrolla RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, pp. 630 y ss.

⁹⁵¹ Esto no impide que en algunos ámbitos existan normas que restrinjan su aplicación a determinadas categorías de daños (por ejemplo, las normas sobre responsabilidad civil por productos defectuosos).

⁹⁵² ÁLVAREZ OLALLA, M.P., *Manual de Derecho de Daños*, Navarra, 2021, p. 19.

del comportamiento de un sistema inteligente o de cualquier otro producto, tecnología, actividad o conducta humana, en la medida en que estos se indemnizarán (o no) bajo los mismos parámetros.

Como viene manteniendo la doctrina, los daños que genera esta tecnología no son daños nuevos o de diferente naturaleza a los que tradicionalmente se vienen distinguiendo en el Derecho de daños (patrimoniales, extrapatrimoniales, personales...) ⁹⁵³. Cuestión distinta será que factores como la escala y la velocidad con la que actúan estos sistemas, así como la ingente cantidad de datos de los que depende su aprendizaje, puedan incrementar su frecuencia, alcance o gravedad.

En definitiva, los sistemas de IA pueden causar los mismos *tipos de daños* a las personas, si bien a través de mecanismos diferentes y con una proyección mucho mayor ya que se extenderán a múltiples sujetos por la interconectividad de esta tecnología. Así, si un sistema de IA utilizado en Bolsa arroja una mala decisión debido a un error en su modelo predictivo y provoca pérdidas millonarias a los inversores, el tipo de perjuicio ocasionado (daño patrimonial) es bien conocido en el Derecho de daños. La particularidad reside en que, como dicho algoritmo de *trading* de alta frecuencia realiza miles de operaciones por segundo, un error podría desencadenar una crisis bursátil en cuestión de minutos. De la misma forma, si un algoritmo de contratación discrimina sistemáticamente a las mujeres para ciertos puestos de trabajo, puede causar un daño moral derivado de la lesión al derecho a la igualdad, siendo también una categoría de daños reconocida en nuestra legislación. Ahora bien, la cantidad de personas afectadas por esa decisión algorítmica discriminatoria será mucho mayor, teniendo en cuenta que los sistemas de IA automatizan decisiones que antes eran individuales y además lo hacen en fracciones de segundo. A lo que debe añadirse que, al «alimentarse» estos sistemas con enormes cantidades de datos históricos, si los mismos están sesgados o son

⁹⁵³ De hecho, como explica RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 635, todas las clasificaciones de los daños que se vienen utilizando se pueden presentar en los daños ocasionados por sistemas de IA. Cuestión distinta será, como así viene ocurriendo, que la imputación objetiva y subjetiva del daño sea más complicada en los casos en que intervenga un sistema de IA que cuando no se emplea dicha tecnología. Pero en opinión de este autor, la indemnización por muerte de un peatón no debería revestir ninguna especialidad por haber sido ocasionada por un vehículo autónomo, la indemnización por la pérdida de oportunidad ocasionada por un mal diagnóstico médico tampoco habría de diferir en caso de haber sido realizado por un algoritmo, ni el daño moral derivado de una vulneración al derecho al honor debería ser distinto si el mensaje injurioso fue obra de una inteligencia artificial generativa. A lo que podría añadirse, respecto a los casos objeto de nuestro estudio y siguiendo este razonamiento, que la reparación por discriminar a una mujer en la concesión o denegación de un préstamo que le impida o dificulte el acceso a una vivienda habría de ser la misma si la decisión la adopta un empleado concreto de la entidad prestamista, o un sistema de *scoring* crediticio plenamente automatizado utilizado por esa entidad.

incorrectos, replicarán y amplificarán dichos errores pudiendo vulnerar los derechos fundamentales de innumerables personas y dando lugar al fenómeno conocido como la «automatización de la discriminación».

Por todo ello, se han identificado una serie de rasgos o cualidades atribuibles a los daños y perjuicios que estos sistemas pueden conllevar: por un lado, su *carácter masivo*, ya que no suelen afectar a un solo sujeto sino que lo normal es que un mismo sistema afecte a una pluralidad de personas o a colectivos; por otro lado, su *naturaleza cumulativa*, ya que el impacto que pueden tener en un sujeto concreto quizá sea mínimo o no significativo, pero la escalabilidad, repetición y amplificación que permite esta tecnología incrementa considerablemente el alcance y gravedad de los daños; y por último, la posibilidad de que se *proyecten reticularmente*, esto es, que aprovechando los efectos de red afecten a sujetos diferentes a la víctima inicial⁹⁵⁴.

No obstante, las particularidades inherentes a estos daños como su impacto estructural y colectivo, el hecho de que puedan ser mínimos o que sus efectos no sean inmediatos, ponen en tela de juicio su compensación. Tal como mantiene RUBÍ PUIG, «abrir la puerta a la compensación de daños difusos, inciertos o difícilmente verificables puede frenar el desarrollo de esta tecnología porque los desarrolladores de sistemas de inteligencia artificial se exponen a contingencias de responsabilidad civil inciertas y, en el límite, a un riesgo de insolvencia»⁹⁵⁵.

2.1.1. Daños indemnizables en la Propuesta de Reglamento del PE de 20 de octubre de 2020, en la Propuesta de Directiva de la CE de 28 de septiembre de 2022 sobre responsabilidad civil en materia de IA y en la Directiva 2024/2853 sobre responsabilidad por productos defectuosos: especial referencia al daño moral

A pesar de la relevancia práctica del tema de los daños indemnizables, la PDRC no le presta especial atención a este requisito y lo contempla sólo de forma incidental al señalar que «*debe exigirse al demandante que demuestre que la información de salida o la no producción de información de salida dio lugar al daño*»⁹⁵⁶, previsión que a nuestro modo de ver no sólo exige la acreditación de la casualidad jurídica sino que también implica, siquiera implícitamente, la prueba del propio daño sufrido por la víctima⁹⁵⁷.

⁹⁵⁴ RUBÍ PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 637.

⁹⁵⁵ RUBÍ PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 647.

⁹⁵⁶ Considerando 22 y art. 4.1. c)

⁹⁵⁷ En mismo sentido se posiciona RUBÍ PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 674.

Ésta es la única previsión normativa sobre el daño que contiene la PDRC. No concreta cuestiones como el *tipo de daños resarcibles* o el *alcance y límites a las cuantías indemnizatorias*, sino que deja su determinación a los Derechos nacionales para garantizar, según mantiene la Comisión, el principio de proporcionalidad⁹⁵⁸. Con todo, su pretensión es que las víctimas puedan reclamar la indemnización de cualquier tipo de daños. Así lo establece al delimitar el ámbito de aplicación de la norma y deslindarlo del de la PDPD, afirmando que «[1]a presente propuesta cubre las demandas nacionales de responsabilidad fundamentadas principalmente en la culpa de cualquier persona con el fin de indemnizar por *cualquier tipo de daño* y a cualquier tipo de víctima».

Se trata de una diferencia importante respecto al régimen de responsabilidad que configuró el Parlamento en 2020, donde sí se recogían normas sobre las *partidas indemnizatorias* exigibles (art.3), el *quantum* de la indemnización (arts. 5 y 9) o su alcance (art. 6), entre otras. Según su art. 2.1, el Reglamento se aplicaría en aquellos casos en que una actividad física o virtual, un dispositivo o un proceso gobernado por un sistema de IA hubiera causado daños o perjuicios a *la vida, la salud, la integridad física de una persona física y los bienes de una persona física o jurídica, o hubiera causado daños morales considerables que dieran lugar a una pérdida económica comprobable*. Por ende, no «cualquier tipo de daño» podía indemnizarse en base a sus reglas, sino que sólo se configuraban como daños indemnizables los corporales, materiales y morales. Estos últimos, además, sólo cuando fueran *significativos y resultaran en una pérdida económica comprobable*⁹⁵⁹.

⁹⁵⁸ Esta opción legislativa se justifica en el Considerando 10 donde se pone de manifiesto que, para garantizar el principio de proporcionalidad, no se deben armonizar «los aspectos generales de la responsabilidad civil que estén regulados de diferentes maneras por las normas nacionales de responsabilidad civil, como [...] los diferentes tipos de daños que dan lugar a demandas por daños y perjuicios [...], el cálculo de los daños y perjuicios [...]», confirmando el Considerando 22 con claridad meridiana que la Directiva «no armoniza las condiciones relacionadas con los daños, por ejemplo, la cuestión de qué daños son indemnizables, que también están regulados por la legislación nacional y de la Unión aplicable».

⁹⁵⁹ El art. 3 i) definía el daño o perjuicio como el «impacto adverso que afecta a la vida, la salud, la integridad física de una persona física, los bienes de una persona física o jurídica o bien que produce daños morales significativos que resultan en una pérdida económica comprobable». Explicaba el Considerando 16 que «El presente Reglamento debe cubrir los daños o perjuicios a la vida, la salud, la integridad física, la propiedad y los daños morales significativos que den lugar a una pérdida económica comprobable por encima de un umbral, armonizado en el Derecho de la Unión en materia de responsabilidad civil, que equilibre el acceso a la justicia de las personas afectadas con los intereses de otras personas implicadas. La Comisión debe reevaluar y ajustar en el Derecho de la Unión los umbrales relativos a los perjuicios. Por *daño moral significativo* debe entenderse un daño como consecuencia del cual la persona afectada sufre un perjuicio considerable, un menoscabo objetivo y demostrable de sus intereses personales y una pérdida económica calculada teniendo en cuenta, por ejemplo, las cifras medias anuales de ingresos anteriores y otras circunstancias pertinentes».

Respecto a los primeros, se incluían tanto los daños físicos que conllevaran la muerte de la persona como los que sólo comportaran lesiones sin ocasionarle la muerte. Estos daños se indemnizaban hasta un máximo de 2 millones de euros. Ahora bien, lo que realmente se indemnizaba no era el daño corporal en sí mismo, sino más bien el daño emergente (costes médicos) o el lucro cesante (ingresos dejados de percibir) como consecuencia del daño corporal, es decir, el perjuicio económico derivado del mismo⁹⁶⁰. También quedaban cubiertos en dicha norma los gastos funerarios y los derivados del cumplimiento de los deberes de asistencia (también a concebidos no nacidos) que estuvieran a cargo del fallecido (pensión similar a la que hubiera tenido que pagar el fallecido, calculada conforme a la esperanza de vida del pensionista)⁹⁶¹.

En cuanto a los daños morales, categoría que adquiere especial relevancia en los casos de discriminación algorítmica como después de analizará, el Reglamento establecía que se indemnizarían los «daños morales significativos» que resultaran en una pérdida económica comprobable o en daños materiales, también cuando distintos bienes de la persona afectada resultaran dañados como resultado de un único funcionamiento de un sistema IA de alto riesgo (art. 5.1 b). Se trataba, según el Considerando 19, de incluir los daños morales significativos siempre y cuando dieran lugar a una pérdida *por encima de un umbral, armonizado en el Derecho de la Unión en materia de responsabilidad civil*, con el fin de equilibrar el acceso a la justicia de los afectados y los intereses de otras personas implicadas.

Por tanto, bajo el paraguas de aquella pretendida regulación quedaban incluidos dentro del sistema de responsabilidad objetiva establecido para los sistemas de *alto riesgo*, todas las repercusiones económicas que se generasen como consecuencia, por ejemplo, del resultado discriminatorio del sistema de IA, como las pérdidas o ganancias dejadas de obtener por la lógica algorítmica, pero también todos los perjuicios económicos derivados del daño moral significativo ocasionado por la acción discriminatoria⁹⁶². No así, el daño moral autónomo⁹⁶³. En definitiva, la norma se refería a un daño moral con repercu-

⁹⁶⁰ Así lo apuntaba ÁLVAREZ OLALLA, M.P., «Propuesta de Reglamento en materia de responsabilidad civil ...», *cit.*, p. 6.

⁹⁶¹ *Ibidem*. Considera esta autora que ni el daño corporal sufrido por la víctima antes del fallecimiento, ni el daño moral de los familiares del fallecido eran indemnizables, teniendo en cuenta que la indemnización se calculaba en función de los costes o gastos derivados y de los perjuicios económicos indirectos.

⁹⁶² En los casos de daños ocasionados por sistemas de no alto riesgo, la Propuesta de Reglamento no regulaba ni el alcance ni el quantum indemnizatorio, sino que remitía al ordenamiento nacional del Estado miembro donde se hubiera producido el daño o perjuicio (vid., art.9).

⁹⁶³ Como apunta NAVAS NAVARRO, S., *Daños ocasionados por sistemas de inteligencia artificial...*, *cit.*, p. 63, esta norma no es extraña en el ámbito de la responsabilidad, ya que, por ejemplo en la

sión económica (*sui generis*), ya que lo que se indemnizaban eran los gastos de tratamiento médico (daño económico emergente) y el perjuicio económico por el cese, suspensión temporal o reducción de la capacidad para generar ingresos (lucro cesante), o el aumento de necesidades acreditado mediante certificado médico⁹⁶⁴.

Esta concepción del daño moral indemnizable fue muy criticada por la doctrina en tanto que, con frecuencia, los daños que pueden ocasionar los sistemas de IA son precisamente daños morales puros o autónomos. No hay más que pensar en los casos objeto de nuestro estudio donde las decisiones discriminatorias adoptadas por estos sistemas, basadas también en muchos casos en perfilados sesgados, se traducen en una afectación a la dignidad de la persona o incluso en pérdidas de oportunidades sin necesidad de que exista un daño material o patrimonial concomitante. En aquel texto el legislador se centró en el resarcimiento de los daños corporales y olvidó la capacidad de los sistemas de IA para producir daños más allá de la esfera patrimonial y afectar a bienes jurídicos distintos a la salud o integridad física, postura que resultaba especialmente perjudicial en los casos de discriminación algorítmica al excluir el resarcimiento de sus daños prototípicos. En base a ello, y también para aumentar la coherencia con el RIA que se publicó unos meses después (el cual sí tenía en cuenta tales riesgos y daños), la doctrina mantuvo la conveniencia de que se incluyeran entre los daños indemnizables tanto los daños morales autónomamente considerados, como los patrimoniales puros⁹⁶⁵.

Ahora bien, la exclusión de tales daños no significaba que no pudieran ser compensados, sino sólo que no lo podrían ser en base al régimen de responsabilidad objetivo previsto para los sistemas de alto riesgo en aquel Reglamento, quedando la víctima abocada a ejercitar diferentes pretensiones, basadas en daños diferentes y sobre la base de legislaciones diferentes (art.2.3), lo cual sin duda suponía un inconveniente considerable⁹⁶⁶.

Similares consideraciones se hicieron respecto a los daños resarcibles bajo la inicial PDPD, que parecía excluir también la compensación de los daños morales. Por un lado, porque en su art. 4.6 definía el daño como «las pérdidas materiales» que resultasen de la lesión de una serie de intereses pro-

responsabilidad del fabricante por productos defectuosos no se contempla, en principio, la indemnización del daño moral autónomo.

⁹⁶⁴ ÁLVAREZ OLALLA, M.P., «Propuesta de Reglamento en materia de responsabilidad civil...», *cit.*, pp. 6-7.

⁹⁶⁵ LLORENS ESPADA, J., *op.cit.*, p. 598. NAVAS NAVARRO, S., *Daños ocasionados por sistemas de inteligencia artificial...*, *cit.*, p. 63.

⁹⁶⁶ En los casos objeto de nuestro estudio la reclamación de los daños morales en sentido estricto debía hacerse al amparo del derecho antidiscriminatorio.

tegidos que se enumeraban⁹⁶⁷. Por otro lado, porque en su articulado no había una disposición como la contenida en el art. 9 de la todavía vigente Directiva de productos defectuosos de 1985, según el cual, «[e]l presente artículo no obstará a las disposiciones nacionales relativas a los daños inmateriales». En consecuencia, la doctrina se planteó si la reclamación de los perjuicios extrapatrimoniales o morales no podía llevarse a cabo en el marco de esta Directiva pudiendo obtenerse su resarcimiento sólo cuando la reclamación se fundamentase en reglas nacionales generales como, por ejemplo, las del art. 1902 CC y concordantes, o en una normativa sectorial interna⁹⁶⁸. Para MARTÍN CASALS, teniendo en cuenta el tenor de su Considerando 18, según el cual «la presente Directiva no deb[ía] afectar a las normas nacionales relativas a los daños morales», aunque la Propuesta sólo hiciera referencia a los daños o perjuicios «patrimoniales» que derivasen de la lesión de una serie de intereses jurídicos (es decir, “muerte o lesiones corporales –incluidos los daños para la salud psicológica comprobados médicamente⁹⁶⁹–”, “daños materiales o a las cosas” y “daños a datos y contenidos digitales distintos a los incluidos en el sistema”), también podría aplicarse a los daños morales derivados de dicha lesión, cuyo resarcimiento sería posible cuando lo permitiera el Derecho interno.

Otros autores se pronunciaron simplemente sobre los daños morales derivados de las lesiones corporales, los cuales entendían cubiertos por la PDPD. Así, para GÓMEZ LIGUERRE, los daños morales podían tener cabida en la reclamación siempre y cuando pudieran ser considerados como «daños para la salud psicológica comprobados médicamente»⁹⁷⁰. Por su parte, RUBÍ PUIG se

⁹⁶⁷ Para RUBÍ PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, pp. 641-642, es «criticable el intento de proceder con una definición de daños y, sobre todo, con la calificación de todos los supuestos previstos como “pérdidas materiales” [...] La terminología ambigua parece dar a entender que solo son resarcibles daños patrimoniales y no los patrimoniales, como particularmente, los daños corporales. Mas la precisión se refiere al carácter cierto y verificable de los daños que tengan un impacto tangible en los intereses protegidos. Sea como sea, el Considerando 17 aclara que las lesiones corporales incluyen los daños para la salud psicológica comprobados médicamente».

⁹⁶⁸ Así lo plantea MARTÍN-CASALS, M., «Las propuestas de la Unión Europea para regular la responsabilidad civil ...», *cit.*, p. 86.

⁹⁶⁹ Para MARTÍN-CASALS, M., «Las propuestas de la Unión Europea para regular la responsabilidad civil...», *cit.*, p. 86, ese añadido «no supone ninguna ampliación del ámbito de protección, sino tan solo una clarificación de la regulación actual, ya que el término “lesiones corporales” engloba tanto los daños físicos como los psíquicos».

⁹⁷⁰ GÓMEZ LIGUERRE, C.I., «La Propuesta de Directiva sobre responsabilidad por daños causados por productos defectuosos», *Indret*, n. 4, 2022, p. 4. El ELI (*European Law Institute*), ya consideraba necesario incluir los daños morales derivados de lesiones corporales bajo el paraguas de la DPD en su *Response of the European Law Institute to the Public Consultation of the European Commission on Civil Liability. Adapting liability rules to the digital age and artificial intelligence*, 2022, p.16. No obstante, mantenía también que, debido a las disparidades en los ordenamientos nacionales acerca del resarcimiento de los daños morales puros o autónomos o la valoración de la indemnización por estos daños, era conveniente abstenerse de regularlos y que fueran los Estados miembros los que adoptasen reglas específicas sobre ellos.

pronuncia sobre los daños derivados de la pérdida o corrupción de datos que no se utilicen exclusivamente con fines profesionales, entendiendo que sólo serán indemnizables los daños patrimoniales, pero no los morales, y criticando dicha previsión por superficial y con escasa aplicación práctica debido a que «los datos cuya pérdida o corrupción pueden tener un valor patrimonial serán las más de las veces datos utilizados en el marco de actividades económicas profesionales [excluidos por el precepto], y los datos utilizados para fines privados darán lugar únicamente a daños morales [también excluidos]»⁹⁷¹.

La versión final de la Directiva 2024/2853 esclarece las dudas sobre esta cuestión. A pesar de que su art. 6.2 sigue haciendo referencia a que el derecho a indemnización cubrirá todas las «*pérdidas materiales*» derivadas de los daños a que se refiere el apartado primero, se añaden dos disposiciones más en el propio articulado: la primera, estableciendo que el derecho a indemnización «cubrirá también los daños morales derivados de los daños a que se refiere el apartado 1, en la medida en que puedan ser indemnizados con arreglo al Derecho nacional» (art. 6.2 *in fine*)⁹⁷²; y la segunda, disponiendo que lo previsto en el precepto no afectará al Derecho nacional relativo a las indemnizaciones en virtud de otros regímenes de responsabilidad (art.6.3).

Además, según establece su art. 2.4.b), la presente Directiva no afectará a «los derechos que la persona perjudicada tenga en virtud de las normas nacionales en materia de responsabilidad contractual o de responsabilidad extracontractual por motivos distintos del carácter defectuoso de un producto con arreglo a lo dispuesto en la presente Directiva, incluidas las normas nacionales por las que se transpone el Derecho de la Unión».

Sea como fuere, sobre lo que parece haber consenso doctrinal es acerca de que los *daños morales autónomos* y los *patrimoniales puros* no están cubiertos por esta Directiva⁹⁷³. No obstante, su resarcimiento podría reclamarse con-

Apunta MARTIN CASALS una cuestión relevante cual es la de saber si estamos ante un daño psíquico que constituye un daño corporal y no ante un simple daño emocional o daño moral puro. Explica que para dicha determinación puede ser fundamental atender a si la perturbación se halla reconocida como enfermedad en la CIE-10 (OMS, 1992)⁵⁵ o en el DSM-5 (2014)⁵⁶, que son dos sistemas clasificatorios ampliamente reconocidos y utilizados en el ámbito internacional para el diagnóstico de los trastornos mentales de la edad adulta y también de la niñez y adolescencia. MARTÍN-CASALS, M., «Las propuestas de la Unión Europea para regular la responsabilidad civil...», *cit.*, p. 87.

⁹⁷¹ RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, 643. Los corchetes son nuestros.

⁹⁷² En el Considerando 23 se reitera que «[...] debe preverse una indemnización por los daños morales resultantes de los daños cubiertos por la presente Directiva, como el dolor y el sufrimiento, en la medida en que dichas pérdidas puedan ser indemnizadas con arreglo al Derecho nacional».

⁹⁷³ Se pregunta GÓMEZ LIGÜERRE, C.I., «La Propuesta de Directiva sobre responsabilidad...», *cit.*, p. 4, por qué no se ha aprovechado la ocasión para contemplar también el daño puramente económico en esta Directiva.

forme al régimen general de la responsabilidad civil o a un régimen especial nacional, si concurren los requisitos para ello. Además, como ya dijimos, cualquier daño resarcible en virtud de un régimen interno de responsabilidad subjetiva o por culpa podría beneficiarse de las reglas armonizadas que contiene la PDRC, si llegara a aprobarse.

Por lo que se refiere en concreto a los daños derivados de la discriminación (algorítmica o no), habrá que acudir a las normas internas del Derecho antidiscriminatorio para determinar *qué tipo de daños* pueden ser resarcidos por la vulneración de este derecho fundamental, así como su *cuantificación y compensación*. Estas normas son el resultado de la transposición al ordenamiento español del acervo comunitario en materia de igualdad. Sólo en aquellas cuestiones no reguladas por esta legislación especial, serán aplicables las normas generales en materia de responsabilidad civil para el resarcimiento de los daños y perjuicios sufridos. A estas cuestiones dedicaremos el siguiente apartado.

2.2. EN PARTICULAR, LOS DAÑOS DERIVADOS DE LA DISCRIMINACIÓN

Las instituciones europeas no vacilan al reconocer que en el escenario de la IA la *discriminación o estigmatización* puede ocasionar daños jurídicamente relevantes. Una consagración expresa de esta realidad la recogía el art. 4 del Reglamento anexo a la Resolución del Parlamento sobre aspectos éticos de la IA, de 20 de octubre de 2020, que disponía en su apartado n) que por «lesión o daño» debía entenderse *toda lesión física o mental y todo daño material o inmaterial, incluso cuando hubieran sido causados por incitación al odio, sesgo, discriminación o estigmatización*. Entre tales daños se mencionaban, a modo de ejemplo, las pérdidas financieras o económicas, pérdida de empleo o de oportunidades educativas, restricción indebida de la libertad de elección o de expresión, pérdida de la privacidad y cualquier otra infracción del Derecho de la Unión que fuese perjudicial para una persona. También en su Resolución sobre un régimen de responsabilidad civil en materia de IA, de la misma fecha, el Parlamento establecía como posible daño ocasionado por los sistemas de IA «la toma de decisiones sesgadas en materia de seguro de enfermedad, concesión de crédito, autos judiciales, contratación o empleo [...]», daño que calificaba de «grave» al poner en peligro la dignidad humana y los valores y libertades europeos.

Que la discriminación, algorítmica o no, puede generar daños resarcibles no es algo nuevo. De hecho, el acervo comunitario en materia de igualdad persigue, desde sus orígenes, la prevención y reparación de los daños que ocasiona la vulneración del derecho a la igualdad y la prohibición de discriminar.

Lo relevante en esta materia es que se exceptúa la regla general del Derecho de daños en virtud de la cual el daño no se presume y es el demandante el que debe acreditar su existencia, mejorándose claramente con ello la posición de los perjudicados.

2.2.1. La presunción del art. 27 de la Ley 15/2022 integral de igualdad

En virtud del art. 27.1 de la Ley 15/2022 integral de igualdad, «[a]creditada la discriminación *se presumirá la existencia de daño moral*». Se trata de una *presunción legal* de existencia de daños, concretamente de daños morales, que recuerda a la que opera en los casos de intromisiones ilegítimas en el derecho al honor, intimidad e imagen de las personas, reconocida en el art.9.3 de la LO 1/1982⁹⁷⁴. En base a dicha norma, no cabe duda de que la vulneración del derecho a la igualdad y no discriminación da derecho a la víctima a reclamar su resarcimiento a través de una indemnización que engloba el daño moral, el cual se presume existente cuando queda acreditada la discriminación injusta. Así, parece que el legislador está reconociendo que, *en todo caso*, ser discriminado ocasiona un daño *per se*, ventaja nada desdeñable para los demandantes⁹⁷⁵.

Cabe plantearse si el demandado podría refutar o no dicha presunción a la vista de que la norma no precisa su carácter *iuris tantum* o *iure et de iure*. Respecto a la disposición del art. 9.3 LO 1/1982, análoga a la aquí analizada, el TS ha mantenido, por un lado, que la misma es *iure et de iure*, y por otro lado, que la *indemnización resultante no puede ser meramente simbólica*, porque de ser así tendría un efecto disuasorio inverso⁹⁷⁶. En base a esta postura jurisprudencial,

⁹⁷⁴ De hecho, como señala PEÑA LÓPEZ, F., «¿Qué daño moral se presume que se deriva de los actos de discriminación y qué tipo de presunción establece la Ley 15/2022?», *CESCO*, septiembre 2022, p.2, «las diferencias entre la norma transcrita en el párrafo anterior y el art. 9.3 de la LO 1/1982 son pocas. En realidad, sólo dos. La primera consiste en que, en la norma más antigua, lo que la ley presume, una vez acreditada la vulneración objetiva del derecho, es “la existencia de perjuicio”, mientras que en la nueva regla lo presumido es “la existencia de daño moral”. La segunda y última diferencia radica en que la nueva ley añade al elenco de criterios de valoración del daño moral que ya estaban en la regla de la LO 1/1982 vigente, uno nuevo: “la concurrencia o interacción de varias causas previstas en la ley”». Para este autor, cuya apreciación compartimos, la relevancia práctica de ambas diferencias es nula o casi nula.

⁹⁷⁵ Para AGUILERA RULL, A., «Daño moral por discriminación», en Gómez Pomar, F./Marín García, I. (coords.) *El daño moral y su cuantificación*, Madrid, 2023, p. 571, el hecho de que en casos de discriminación la indemnización sea automática pretende facilitar la reclamación de los daños ocasionados y es reflejo de la relevancia del bien jurídico protegido.

⁹⁷⁶ Entre otras, la STS 267/2023, de 20 de febrero (ECLI:ES:TS:2023:989), la STS 130/2020, de 27 de febrero (RJ 2020/613) y la STS 512/2017, de 21 de septiembre (ECLI:ES:TS:2017:3322).

la doctrina augura que con toda probabilidad se mantendrá el mismo criterio respecto a la presunción del art. 27.1 *in fine*, de forma que no se permitirá al demandado presentar prueba en contrario, esto es, probar que aun existiendo discriminación la misma no ocasionó daños a la víctima⁹⁷⁷.

No obstante, un sector doctrinal aboga por el carácter *iuris tantum* de la presunción para mantener la coherencia con la función reparadora del Derecho de daños y con el principio de reparación integral del daño. Dicho principio no solo impide que se indemnice más allá de los daños probados, sino también que se reparen daños inexistentes. Por ello, en opinión de estos autores, presumir sin posibilidad de prueba en contrario que de la discriminación se irrogan siempre y en todo caso daños para la persona o grupo discriminado, llevará a reparar daños que en ocasiones podrían no existir, en clara contradicción con dicho principio⁹⁷⁸. Todo ello sin perjuicio de que, además,

Reconociendo también el carácter *iure et de iure* de la presunción del art. 9.3 LO 1/1982, las SSTS 312/2014, de 5 de junio (RJ 2014/3087), 261/2017, de 26 de abril, 388/2018 de 21 de junio (RJ 2018/2771). En esta última se afirma con claridad, por un lado, el carácter *iure et de iure* de dicha presunción de daños (citando las sentencias núm. 964/2000, de 19 de octubre, y núm. 12/2014, de 22 de enero); y por otro lado, citando numerosas sentencias de la Sala Primera, y concretamente la núm. 386/2011, de 12 de diciembre, que la indemnización no puede ser simbólica: «según la jurisprudencia de esta Sala (SSTS de 18 de noviembre de 2002 y 28 de abril de 2003) no es admisible que se fijen indemnizaciones de carácter simbólico, pues al tratarse de derechos protegidos por la CE como derechos reales y efectivos, con la indemnización solicitada se convierte la garantía jurisdiccional en un acto meramente ritual o simbólico incompatible con el contenido de los artículos 9.1, 1.1. y 53.2 CE y la correlativa exigencia de una reparación acorde con el relieve de los valores e intereses en juego (STC 186/2001, FJ 8)» (STS 4 de diciembre 2014, rec. núm. 810/2013)». En contadas ocasiones el TS ha calificado esta presunción como *iuris tantum*. Ejemplo de esto lo encontramos en la STS 604/2011, de 20 julio 2011 (RJ 2011/6126). En otras ocasiones, mantiene que esta presunción es «una realidad *“in re ipsa”*» que impide cualquier prueba en contra (por ejemplo, en las SSTS núm. 94/2009, 25 febrero 2009 (RJ 2009/1515), núm. 284/2011 de 11 abril 2011 (RJ 2011/3446), o de 27 octubre 2012 (RJ 2012/1129): «La norma legal establece que la existencia del perjuicio se presumirá siempre que se acredite la intromisión ilegítima (art. 9.3, inciso primero, LO). No se trata de una mera presunción *“iuris tantum”*, sino que la intromisión ilícita supone la existencia del perjuicio indemnizable, a modo de una realidad *“in re ipsa”*. El perjuicio presumido es el correspondiente al daño moral, que no obsta a la prueba de un daño material, y para cuya cuantificación económica, la norma legal del art. 9.3 indicado, señala unas pautas [...]».

⁹⁷⁷ Así lo entiende PEÑA LÓPEZ, F., «¿Qué daño moral se presume ...», *cit.*, p. 4; CASADESÚS RIPOLL, P., *op.cit.*, p. 96. Para ARIAS DOMÍNGUEZ, A., *La cuantificación de la indemnización por daño moral por trasgresión de derechos fundamentales en los despidos nulos*, Colección de Derecho del Trabajo y Seguridad Social, n° 15, Madrid, 2023, p. 2, no hay duda de que la norma entiende que siempre que haya discriminación se causa un quebranto moral, y «aunque parece estar configurando una presunción *iuris tantum*, la mecánica a la que nos ha arrastrado la interpretación del artículo 183 LRJS nos avocará a entender que estamos en presencia de una hermenéutica que no admite excepciones a la regla general».

⁹⁷⁸ En este sentido, PEÑA LÓPEZ, F., «¿Qué daño moral se presume...», *cit.*, p. 5. Considera este autor que, si bien en la mayoría de ocasiones los actos de discriminación ocasionarán daños morales puros, esto no sucederá siempre. Además, para PEÑA, esta indemnización automática es más una especie de multa que una verdadera indemnización. También RUBI PUIG, A., «Inteligencia arti-

dicha interpretación se aparta de la regla general del art. 385.3 LEC, según la cual, «[l]as presunciones establecidas por la ley admitirán la prueba en contrario, salvo en los casos en que aquélla expresamente lo prohíba». Por tanto, a falta de prohibición expresa en tal sentido, debería entenderse que la misma es *iuris tantum*.

Como afirma PEÑA LÓPEZ, mantener el carácter *iuris tantum* de la presunción tendría una doble ventaja: por un lado, constituiría un incentivo para que las partes proporcionaran al juez una mayor información y pruebas sobre los efectos emocionales de la conducta, lo que situaría a éste en mejor posición para valorar los perjuicios morales puros sufridos por la víctima; y por otro lado, evitaría que en estos casos se desvirtuara la función reparadora de la responsabilidad civil pasando a convertirse en una especie de sanción punitiva⁹⁷⁹. Además, tampoco supondría una gran ventaja para los discriminadores teniendo en cuenta que en circunstancias normales las víctimas se beneficiarían de la indemnización, ya que el hecho mismo de establecer una presunción a su favor refleja que generalmente la discriminación provoca daños morales⁹⁸⁰.

Otra opción interpretativa que propone el autor citado, en virtud de la cual podría explicarse el carácter *iure et de iure* de la presunción analizada, sería la de mantener que los daños morales derivados de la discriminación son una especie de «daños en sí», es decir, el tipo de daño que el DCFR denomina «*injury as such*», que se suele vincular a la lesión de los derechos de la personalidad y consiste en la propia lesión del derecho en sí mismo. Si fuese así, cobraría sentido que la presunción del art. 27 Ley 15/2022 fuera *iure et de iure*, ya que entonces la discriminación y el «daño en sí» vendrían a coincidir. Esta conceptualización del daño implicaría dos cosas: por un lado, supondría un reconocimiento de que para el legislador el hecho de discriminar constituye un daño para todos los discriminados (sin perjuicio de que también les pudiera producir zozobra, ansiedad, dolor o tristeza y de que les hubiera ocasionado además daños económicos); y por otro lado, encajaría con los criterios de valoración establecidos en el precepto que parecen ir dirigidos a medir la intensidad de la lesión del derecho y no las emociones o sentimientos negativos de la víctima. Además, teniendo en cuenta que los «daños en sí» poseen naturaleza extrapatrimonial o moral, se entendería la referencia exclusiva a ellos en el precepto comentado⁹⁸¹.

ficial y daños indemnizables», *cit.*, p. 677, siguiendo al primero entiende que esta interpretación de la presunción altera la función compensatoria del derecho de daños y la transforma en una punitiva.

⁹⁷⁹ Para ÁLVAREZ OLALLA, M.P., *Manual de Derecho de Daños*, *cit.*, p. 49, otro caso que guarda cierta similitud con los daños punitivos es la regulación contenida en el art. 9.2 d) de la LO 1/1982.

⁹⁸⁰ PEÑA LÓPEZ, F., «¿Qué daño moral se presume...», *cit.*, pp. 6-7.

⁹⁸¹ Sin perjuicio de que, a juicio de PEÑA, esta posibilidad interpretativa solventaría los problemas de coherencia normativa entre el art. 27 Ley 15/2022 (y el art. 9.3 LO 1/1982) y los princi-

La cuestión de si la mera vulneración de un derecho fundamental da lugar automáticamente a un daño moral ha sido polémica en nuestros tribunales. Si bien la jurisprudencia de la Sala de lo Civil del TS ha mantenido dicha automaticidad desde un primer momento –justificándola en la extrema importancia del bien jurídico protegido y en la dificultad de probar el daño moral–, la Sala de lo Social ha vacilado y cambiado de criterio a lo largo de los años⁹⁸². No obstante, en la actualidad parece aproximarse a la postura mantenida en la jurisdicción civil, ya que, tal como afirma en su sentencia de 2 de febrero de 2015, *la existencia de algunos daños de carácter moral se pone de manifiesto a través de la mera acreditación de la lesión*⁹⁸³. Conforme a tal doctrina, contenida también en su sentencia de 5 de octubre de 2017⁹⁸⁴, *los daños morales resultan indisolublemente unidos a la vulneración del derecho fundamental y, al ser especial-*

prios del Derecho de daños, generaría otros nuevos como el de la consistencia de la propia noción de «daño en sí» o el de la delimitación de su ámbito de aplicación (vid., pp. 7-8).

⁹⁸² Explicando esta evolución CASADESÚS RIPOLL, P., *op.cit.*, pp. 91-94. También CRESPO ORTIZ, D., «La indemnización adicional por vulneración de derechos fundamentales del Artículo 183 LRJS. Reflexiones en torno a sus pretensiones y su cálculo», *Labos*, Vol. 4, No. 3, pp. 157, basándose en la STS de 5 de octubre de 2017 –citada en el texto–, explica que el propio TS reconoce que han sido al menos cuatro las posiciones que se han mantenido sobre si toda vulneración de un derecho fundamental da lugar a un daño moral de forma inherente y sobre cómo proceder a su valoración. No obstante, en opinión de este autor, un análisis más detenido de la cuestión permite afirmar que en realidad no ha acontecido cambio o contradicción alguna a lo largo del tiempo, ya que, el TS en ningún momento ha negado que la vulneración de un derecho fundamental lleva aparejado un daño moral. De hecho, mantiene la automaticidad del daño y su indemnización. Lo único que ha variado en el tiempo son las exigencias para el cálculo de dicho daño hasta el momento actual, donde no se exige que el perjudicado aporte bases o apoyos concretos que permitan arrojar una cifra del daño sufrido. «Se ha llegado a la conclusión de que su apreciación por parte del tribunal, y su cuantificación, se hará de acuerdo con unas bases y apoyos aportados por quien lo reclama. Sin embargo, la aportación de tales indicios no puede ser exigida para los daños morales, y si estas bases o apoyos no son presentados en el juicio, tal omisión no puede conllevar que no se determine indemnización alguna, pues el daño, efectivamente, ha sido causado y debe compensarse, aunque para su cuantificación se aleguen unas cifras genéricas sin colegir del relato fáctico un “sufrimiento personal, daño moral y en la personalidad del actor”, siendo en la práctica tan sólo suficiente la existencia de una vulneración de derechos fundamentales». Para ARIAS DOMÍNGUEZ, A., *op.cit.*, pp.15-16, de la lectura del art. 183 LRJS y preceptos concordantes, se deducen entre otras premisas las siguientes: por un lado, que si bien no es exigible que la demanda de tutela solicite una cuantía concreta para la indemnización del daño moral, esto no sólo es conveniente sino que el éxito de la pretensión puede depender de esa circunstancia; por otro lado, que se exige que el demandante especifique «los diversos daños y perjuicios» de carácter económico padecidos por la vulneración del derecho fundamental, daños que son distintos a los morales.

⁹⁸³ Vid., STS 809/2015 de 2 de febrero; 201/2017, 19 de diciembre; 480/2019 de 20 de junio 2019.

⁹⁸⁴ Rec. 2497/2015. Es ésta una de las sentencias de referencia en el orden social sobre el daño moral y su cuantificación. Además, remite a otras anteriores también trascendentes: SSTs de 17 diciembre de 2013 (rec. 109/2012), 8 de julio 2014 (rec. 282/2013), 2 de febrero 2015 (rec. 279/2013), 26 de abril 2016 (rec. 113/2015 y 12 de julio 2016 (rec. 361/2014).

mente difícil su estimación detallada, *deben flexibilizarse las exigencias normales para la determinación de la indemnización.*

A nuestro juicio, esta línea jurisprudencial parece chocar con la reciente doctrina sentada por el TJUE en su sentencia de 4 de mayo de 2023, asunto C-300/21 Österreichische Post AG, donde trata concretamente la responsabilidad civil derivada de la infracción del derecho fundamental a la protección de datos personales prevista en el art. 82 RGPD⁹⁸⁵. En síntesis, por lo que ahora interesa, entiende el TJUE que la mera vulneración del derecho a la protección de datos personales no da derecho a una indemnización al perjudicado, independientemente de si se ha producido un daño o no. En otras palabras, la indemnización contemplada en el art. 82 RGPD por la vulneración de este derecho fundamental está condicionada a que el interesado haya sufrido efectivamente daños y perjuicios que deriven precisamente de tal infracción, daños que además deben ser probados y cuantificados conforme a los criterios de cuantificación propios de cada legislación nacional, los cuales habrán de respetar los principios de equivalencia y efectividad del Derecho de la UE. Por tanto, como puede observarse, el TJUE no se muestra partidario de la doctrina de la automaticidad de la indemnización por daño moral derivado de la vulneración de este derecho fundamental en concreto. Es más, nuestro TS tampoco la ha acogido en su reciente sentencia de 19 de marzo de 2024⁹⁸⁶, donde incorpora y aplica los criterios sentados por el tribunal europeo. En concreto, citando expresamente las resoluciones supranacionales de 4 de mayo y 14 de diciembre de 2023, destaca lo siguiente:

En primer lugar, que «no puede considerarse que toda infracción de las disposiciones sobre protección de datos personales dé lugar, *por sí sola*, a un derecho a una indemnización a favor del interesado⁹⁸⁷. Por el contrario, para la pertinencia de la indemnización deberían concurrir tres requisitos cumulativos: (i) un tratamiento de datos personales en infracción de las disposiciones legales pertinentes; (ii) la existencia de daños y perjuicios para el interesado; y (iii) una relación de causalidad entre dicho tratamiento ilícito y esos daños y perjuicios»; y en segundo lugar, «que es preciso que el afectado pruebe que la vulneración de la normativa de protección de datos le haya causado algún perjuicio, por mínimo que sea, y que [...] debe demostrar que las consecuencias de esa infracción que afirma haber sufrido constituyen un perjuicio distinto de la mera infracción de las disposiciones de dicho Reglamento». En definitiva, parece que el TS descarta ahora que la lesión de ese derecho funda-

⁹⁸⁵ Posteriormente y en el mismo sentido, la STJUE (Sala Tercera) de 14 de diciembre de 2023, asunto C456/22, VX y AT contra Gemeinde Ummendorf.

⁹⁸⁶ STS 1495/2024 (ECLI:ES:TS:2024:1495).

⁹⁸⁷ La cursiva es nuestra.

mental lleve indisolublemente unido el reconocimiento del daño moral; por tanto, no bastará con acreditar su vulneración para decretar la indemnización a la víctima.

Algún autor, como ARIAS DOMÍNGUEZ, también precisa que «el automatismo lo es del daño moral una vez constatada la lesión del Derecho Fundamental. Pero dicho automatismo no puede llevarse más allá de lo razonable, deduciendo de una mera alegación de parte de la quiebra de un Derecho Fundamental una indemnización económica por daño moral. Para que exista esa conclusión en relación con la indemnización económica de un daño moral debe existir, de entrada, una actividad procesal de la parte que identifique claramente, en primer lugar, el Derecho Fundamental vulnerado, alegue, en segundo lugar, que dicha quiebra ha lesionado un derecho de orden moral de la persona trabajadora (o que estamos en presencia de una lesión de la que pueda deducirse *prima facie* dicha conclusión), solicite expresamente en tercer lugar, una indemnización económica por el daño moral infligido, pruebe, aunque sea indiciariamente, en cuarto lugar, la existencia así como la entidad de los daños causados y que aporte, en quinto y en último lugar, las bases o elementos claves en que basar la indemnización económica y su proceso de monetización». Para este autor, el art. 183 LRJS, solo libera al perjudicado de la última de las obligaciones enunciadas, por lo que el incumplimiento de las otras provocará un fracaso de la acción interpuesta.

Sea como fuere, de lo que no cabe duda es de que, si se quiere mantener la coherencia entre la normativa antidiscriminatoria y el Derecho de daños, debería abogarse por defender, por un lado, que la existencia del daño es también en este ámbito un presupuesto *sine qua non* para que proceda la indemnización; y por otro lado, que aun cuando la importancia del bien jurídico protegido llevara a presumir la existencia de daños morales en el sujeto discriminado, dicha presunción debería permitir la prueba en contrario por el demandado.

2.2.2. El daño moral presunto y su valoración

El daño moral al que hace referencia el legislador en el art. 27 Ley 15/2022 es el *puro o autónomo*, esto es, el que se identifica con sensaciones y emociones humanas negativas como la zozobra, la ansiedad, el dolor, la inquietud, la tristeza, el miedo, la aflicción o la angustia, provocadas por el evento dañoso⁹⁸⁸.

⁹⁸⁸ En este sentido, PEÑA LÓPEZ, F., «¿Qué daño moral se presume ...», *cit.*, p.3, en base a la interpretación que se ha venido haciendo del art. 9.3 LO 1/1982, norma en la que, en su opinión, se inspira el legislador en el art. 27 Ley 15/2022. La STS de 4 diciembre 2014 (RJ 2014/6360) define

Es decir, se trata de padecimientos que no comportan necesariamente una enfermedad psíquica y que no son médicamente constatables ni evaluables, circunstancia que como se dijo conlleva dificultades en cuanto a su prueba⁹⁸⁹.

Que el legislador se refiera a este tipo de daños es lógico si se tiene en cuenta que en los casos en que se vulneran derechos extrapatrimoniales, los perjuicios de mayor gravedad suelen ser los morales y en muchos casos ni siquiera habrá daños patrimoniales⁹⁹⁰. Por tanto, la víctima podrá reclamar su resarcimiento aun cuando del mismo no deriven perjuicios de carácter económico, esto es, aun cuando no sea un daño moral con repercusión económica, al que se refería el Reglamento propuesto por el Parlamento Europeo en 2020.

Como es sabido, el mayor problema que presenta el daño moral es el de su *valoración o cuantificación* de su importe⁹⁹¹. Más si cabe en el ámbito de la discriminación, ya que, a diferencia de otros daños no patrimoniales con incidencia anatómico-fisiológica, los daños morales puros que derivan de la dis-

claramente el daño moral que se presume: «Provocan daño moral las intromisiones en el honor e intimidad y los ataques al prestigio profesional, tanto más cuando provocan sufrimiento o padecimiento psíquico, que concurre en diversas situaciones como el impacto o sufrimiento psíquico o espiritual, impotencia, zozobra (como sensación anímica de inquietud, pesadumbre, temor o presagio de incertidumbre), ansiedad, angustia, incertidumbre, impacto, quebranto y otras situaciones similares». El TS ha definido el daño moral como «aquel que está representado por el impacto o sufrimiento psíquico o espiritual que en la persona pueden producir ciertas conductas, actividades o, incluso, resultados, tanto si implican una agresión directa o inmediata a bienes materiales, cual si el ataque afecta al acervo extrapatrimonial o de la personalidad (ofensas a la fama, al honor, honestidad, muerte de persona propietario)» (STS 25 de junio de 1984 [RJ 1986/1145]). Más recientemente, la STS 583/2015, de 23 de octubre, entiende como daños morales «aquellos que no son susceptibles de ser evaluados patrimonialmente por consistir en un menoscabo cuya sustancia puede recaer no solo en el ámbito moral estricto, sino también en el ámbito psicofísico de la persona y consiste, paradigmáticamente, en los sufrimientos, padecimientos o menoscabos experimentados que no tienen directa y secuencialmente una traducción económica» y la STS 366/2010, de 15 de junio, afirma que «los daños morales, asociados frecuentemente por la jurisprudencia los padecimientos físicos o psíquicos, son aquellos que afectan a la integridad, la dignidad de la libertad de la persona, como bienes básicos de la personalidad».

⁹⁸⁹ AGUILERA RULL, A., «Daño moral por discriminación», *cit.*, p. 571.

⁹⁹⁰ DÍAZ ALABART, S., «La responsabilidad civil por discriminación ...», *cit.*, para quien, aun cuando las vulneraciones de estos derechos pueden ocasionar daños patrimoniales, afectan principalmente a la esfera personal del individuo.

⁹⁹¹ Como apunta YZQUIERDO TOLSADA, M., *Responsabilidad civil extracontractual. Parte general*, Madrid, 2020, p. 207, no hay que confundir las dificultades que conlleva la prueba del daño moral o corporal, con las dificultades que implica su valoración, puesto que esto último ya es tarea del juzgador. En el mismo sentido, Sobre el tema puede verse actualmente, MARÍN GARCÍA, I./ GÓMEZ POMAR, F. (coords.) *El daño moral y su cuantificación*, Barcelona, 2023.

criminación son mucho más complejos de cuantificar⁹⁹². Además, nuestro ordenamiento jurídico no cuenta con baremos en este ámbito ni establece topes indemnizatorios, con lo que la determinación de la cuantía indemnizatoria quedará al arbitrio de los tribunales en cada caso⁹⁹³.

Sobre esta cuestión es importante la doctrina sentada por la Sala de lo Social del TS en relación con los daños morales derivados de la vulneración de un derecho fundamental. Mantiene el TS que, tal como dispone el art. 179 LRJS, *la demanda deberá expresar con claridad los hechos constitutivos de la vulneración y la cuantía de la indemnización pretendida, en su caso, con la adecuada especificación de los diversos daños y perjuicios, a los efectos de lo dispuesto en los artículos 182 y 183, y que, salvo en el caso de los daños morales unidos a la vulneración del derecho fundamental cuando resulte difícil su estimación detallada, deberá establecer las circunstancias relevantes para la determinación de la indemnización solicitada, incluyendo la gravedad, duración y consecuencias del daño, o las bases de cálculo de los perjuicios estimados para el trabajador*. En virtud de tal doctrina, parece que la regla general sobre la necesidad de acreditación objetiva de la cuantía del daño queda excepcionada en el supuesto de los daños morales unidos a la vulneración del derecho fundamental *cuando resulte difícil su estimación detallada*, ya que en este caso no será necesaria para la víctima cumplir con el requisito de su acreditación⁹⁹⁴.

⁹⁹² De «dificultades extremas» de cuantificación habla AGUILERA RULL, A., «Daño moral por discriminación», *cit.*, p. 571.

⁹⁹³ Como recuerda ITURMENDI MORALES, G., *Responsabilidad Civil por Discriminación*, Madrid, 2023, pp. 14-15, los criterios para la cuantificación de la indemnización son cambiantes y el Tribunal Supremo ha ido fijando las reglas en función de cada jurisdicción sobre la base de la restitución a la víctima en la situación anterior a la producción del daño. En la jurisdicción social, para la cuantificación de los daños morales el TS en numerosas sentencias (entre otras, la STS 214/2022, de 9 de marzo, que cita otras muchas como las SSTs de 15 de febrero de 2012 (Rcud. 6701); de 8 de julio de 2014 (Rec. 282/13; de 2 de febrero de 2015 (Rec. 279/13); de 19 de diciembre de 2017 (Rcud. 624/2016 y de 13 de diciembre de 2018) viene considerando idóneo y razonable la utilización del criterio orientador de las sanciones pecuniarias previstas por la LISOS, que también ha sido admitido por la jurisprudencia constitucional (STC 247/2006, de 24 de julio). No se trata de hacer una aplicación automática y directa del importe de las sanciones, sino que se atiende a la razonabilidad de algunas de esas cifras para la solución del caso, en función de la gravedad de la vulneración del derecho fundamental, buscando no tanto el objetivo propiamente resarcitorio cuanto el aspecto preventivo que ha de corresponder a la indemnización en estos casos (vid., STS 214/2022, de 9 de marzo).

Del estudio jurisprudencial en materia de igualdad y no discriminación llevado a cabo por AGUILERA RULL, la autora extrae dos conclusiones: la primera, que el hecho de no contar con criterios legales de valoración ha llevado a que las indemnizaciones concedidas por los tribunales en los casos analizados sean variables; y en segundo lugar, que algunos tribunales vienen aplicando a los casos de discriminación los baremos del RDL 8/2004, de 29 de octubre, para conseguir proporcionalidad e igualdad en las indemnizaciones. Vid., «Daño moral por discriminación», *cit.*, p. 575.

⁹⁹⁴ Concluye ARIAS DOMÍNGUEZ, A., *op.cit.*, p. 51, que según esta postura, a la hora cuantificar y monetizar el daño moral «no se aplicarán las reglas generales de valoración del daño que se utilizan para la acreditación y cuantificación de los perjuicios de tipo patrimonial».

Además, el art. 183.2 LRJS «*abre la vía a la posibilidad de que sea el órgano judicial el que establezca prudencialmente su cuantía, sin que pueda exigirse al reclamante la aportación de bases más exactas y precisas para su determinación, en tanto que en esta materia se produce la “inexistencia de parámetros que permitan con precisión traducir en términos económicos el sufrimiento que el daño [moral] esencialmente consiste... [lo que] lleva, por una parte, a un mayor margen de discrecionalidad en la valoración... y, por otra parte, “diluye en cierta medida la relevancia para el cálculo del quantum indemnizatorio” de la aplicación de parámetros objetivos, pues “los sufrimientos, padecimientos o menoscabos experimentados” no tienen directa o secuencialmente una traducción económica*» [SSTS/I^a 27/07/06 Ar. 6548; y SSTS/4^a 28/02/08 –rec. 110/01–]” (SSTS 21/09/09 –rcud 2738/08–; y 11/06/12 –rcud 3336/11)”, de tal forma que, en atención a la nueva regulación que se ha producido en la materia tras la LRJS se considera que *la exigible identificación de “circunstancias relevantes para la determinación de la indemnización solicitada” ha de excepcionarse en el caso de los daños morales unidos a la vulneración del derecho fundamental cuando resulte difícil su estimación detallada*»⁹⁹⁵. En definitiva, según el TS, es suficiente a efectos de su reconocimiento por el órgano judicial solicitar en la demanda la indemnización, no siendo necesariamente exigible un mayor grado de concreción en la exposición de parámetros objetivos de muy difícil cumplimiento en atención a la propia naturaleza de los daños morales reclamados.

Con carácter general, los tribunales suelen considerar varios factores para la valoración del daño moral, como la gravedad de la lesión, la duración del sufrimiento, las circunstancias personales de la víctima y el contexto en que ocurrió el daño. No obstante, en materia de igualdad y no discriminación el legislador facilita esta labor proporcionando una serie de parámetros en base a los cuales valorar la intensidad de la lesión, algunos de los cuales recuerdan a los ya establecidos en el art. 9.3 LO 1/1982, como la referencia a la «difusión

⁹⁹⁵ STS 214/2022, de 9 de marzo. En relación con la concreción de la cuantía del daño moral, la STS de 5 de octubre de 2017, anteriormente citada, precisa que cuando la operación que monetiza el daño sea difícil o costosa para el justiciable la determinación debe hacerse por el Juez o Tribunal «prudencialmente», reforzando la exigencia legal. Ahora bien, esa apreciación prudencial solo opera para la cuantificación de la indemnización por daño moral, no extendiéndose la regla para la concreción de otras indemnizaciones, por ejemplo las procedentes de daños y perjuicios de carácter económico derivados de la vulneración del derecho fundamental. También declara la sentencia mencionada que esa indemnización cumple varias finalidades: el restablecimiento de la situación, la reparación de la lesión causada y la prevención de conductas similares, manejando dos parámetros generales de calibración: la suficiencia del monto indemnizatorio, y la integridad de la reparación. Como explica CRESPO ORTIZ, D., *op.cit.*, p. 159, «Lo que pretende la indemnización [del art. 183 LRJS] es compensar el daño moral que indefectiblemente ha producido la vulneración, aunque no sea posible demostrar dicho daño como un sufrimiento concreto, y disuadir de realizar tales comportamientos en un futuro».

o audiencia del medio a través del que se haya producido»⁹⁹⁶. Así, según expresa el art. 27.1 *in fine*, dicho daño «se valorará atendiendo a las *circunstancias del caso*, a la concurrencia o interacción de *varias causas* de discriminación previstas en la ley y a la *gravedad de la lesión* efectivamente producida, para lo que se tendrá en cuenta, en su caso, la *difusión o audiencia del medio a través del que se haya producido*». La referencia a la concurrencia o interacción de varias causas permite deducir que el perjuicio ocasionado a la víctima es mayor cuando se trata de discriminaciones múltiples o intersectoriales en las que confluyen varios factores y que presentan, por ello, un plus de reprochabilidad. Y la gravedad de la lesión sufrida se refiere a las consecuencias de la discriminación y no al hecho generador en sí mismo considerado, ya que el sistema español de responsabilidad civil se basa en la reparación del daño y no en el daño punitivo⁹⁹⁷.

En cualquier caso, como también viene manteniendo nuestro Alto Tribunal, es muy difícil que el daño moral pueda llegar a ser «verdaderamente resarcido», sino que simplemente puede tratar de «compensarse en cierta medida»⁹⁹⁸. Así, en su sentencia de 7 de diciembre de 2006, afirma que ante la «inexistencia de parámetros que permitan con precisión traducir en términos económicos el sufrimiento en que tal daño esencialmente consiste», no se trataría propiamente de un resarcimiento en sentido propio, sino más bien de una vía para «*proporcionar en la medida de lo humanamente posible una satisfacción como compensación al sufrimiento que se ha causado, lo que conlleva la determinación de la cuantía de la indemnización apreciando las circunstancias concurrentes*» (SSTS 25 junio de 1984; 28 de abril de 2005; 10 de febrero de 2006, entre otras)⁹⁹⁹.

Además, junto con el daño moral puro y presunto, también serían resarcibles otros daños o perjuicios ocasionados por la decisión discriminatoria como, por ejemplo, pérdidas patrimoniales o incluso daños corporales, aunque estos últimos serán excepcionales o infrecuentes¹⁰⁰⁰. No obstante, a dife-

⁹⁹⁶ Sobre estos criterios considera ARIAS DOMÍNGUEZ, A., *op.cit.*, p. 23, que si bien es cierto que algunos de ellos se han incorporado «sin demasiada lógica interna, directamente de otras normas civiles», aun así debe aplaudirse la enunciación de parámetros adicionales para valorar el daño moral.

⁹⁹⁷ ITURMENDI MORALES, G., *op.cit.*, p. 15.

⁹⁹⁸ STS 13 de julio de 2015, rec. 221/2014. ECLI ES:TS:2015:4356.

⁹⁹⁹ STS 1303/2006 de 7 de diciembre de 2006 (RJ 2007/266).

¹⁰⁰⁰ Afirma ITURMENDI MORALES, G., *op.cit.*, p. 14, que los daños indemnizables derivados de la discriminación pueden ser *personales y materiales* (tanto de carácter directo como consecutivos o derivados de los previos daños personales y materiales); *patrimoniales puros*, que no requieren de previo daño material o corporal; y finalmente, *daños morales*. Y para AGUILERA RULL, A., «Daño moral por discriminación», *cit.*, p. 557, el daño por discriminación es en gran medida un daño moral y solo excepcionalmente se ocasionarán daños materiales o personales. No obstante, matiza que dependiendo de la conducta discriminatoria concreta la naturaleza de los daños ocasionados podrá variar.

rencia de los primeros, estos deberán probarse puesto que no se presumen *ex lege*.

Todos estos daños podrían darse en situaciones de discriminación, tanto algorítmica como tradicional, sin perjuicio de que el daño moral adquiera una especial preponderancia en este ámbito por la particular lesión a la dignidad de la víctima que supone toda conducta discriminatoria¹⁰⁰¹. Por tanto, podría afirmarse que respecto al *tipo de daños* indemnizables la discriminación algorítmica no presenta especialidades frente a la discriminación que hemos denominado «analógica». Así, el hecho de que la discriminación se haya llevado a cabo a través de sistemas inteligentes, y no al margen de estos, será indiferente por cuanto atañe a la clase de daño que puede sufrir la víctima. Por ejemplo, dará lo mismo que un candidato sea descartado en un proceso de selección laboral por una persona prejuiciosa, que por un *software* programado para hacer un cribado de *curriculums* que presente sesgos en detrimento de las mujeres, por ejemplo. En ambos casos, el daño moral o la pérdida de oportunidad para la solicitante de empleo será idéntico¹⁰⁰².

Por ejemplo, en los casos de discriminación por razón de sexo los daños pueden ser exclusivamente emocionales; sin embargo, en casos de no contrataciones o despidos discriminatorios, los daños podrán ser tanto patrimoniales como morales. También la jurisprudencia del TS avala que en ciertos casos de discriminación laboral, como la retributiva, para la oportuna reparación del derecho fundamental, deben indemnizarse tanto los daños materiales (consistentes, precisamente, en la remuneración dejada de percibir) y los daños morales consustanciales a la violación de cualquier derecho fundamental (vid., STS 43/2017, de 24 de enero). También la reciente STS 524/2024 de 3 de abril, mantiene que «n el fondo de tal postura está la consideración de que en las situaciones en que se invoca la vulneración de un derecho fundamental, como es el de igualdad en su vertiente de derecho a percibir la remuneración correspondiente, «el daño a resarcir no es uno sólo, sino que son dos: a) de un lado, la pérdida de una parte del salario que ha de atribuirse al incumplimiento empresarial y que tiene una indemnización legalmente tasada, esto es, la remuneración prevista normativamente; y b) de otro, el daño moral que ha de producir –en términos generales– esa conculcación del derecho fundamental y que forzosamente ha de imputarse al infractor, a quien –además– le es exigible por tal consecuencia la indemnización prevista en el art. 1101 CC».

¹⁰⁰¹ En este sentido GARCÍA RUBIO, M.P., «Discriminación por razón de sexo ...», *cit.*, p. 163, manifiesta con rotundidad que, en casos de discriminación, son indemnizables tanto los daños patrimoniales como los daños morales o extrapatrimoniales. En el ámbito de la IA, RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 648, reconoce que «es esperable que muchos de los perjuicios que generen [los sistemas de IA] consistan en un daño moral puro» y cita precisamente como ejemplo el caso que nos ocupa en este trabajo, esto es, los sesgos discriminatorios de un algoritmo que, en muchos casos, sólo implicarán la vulneración de derecho a la igualdad de una persona, pero no le comportarán un daño de tipo patrimonial. También CASADESÚS RIPOLL, P., *op.cit.*, p. 91 y ss.

¹⁰⁰² Debe apuntarse que, si bien la teoría del daño por pérdida de oportunidad carece de sustento normativo en nuestro ordenamiento jurídico, ha sido acogida por la doctrina y la jurisprudencia. Por ello, no vemos inconveniente en que se aplique a los casos en que la decisión adoptada por un sistema de IA discriminatorio conlleva dicha pérdida de oportunidad y el daño consiguiente. Sobre la pérdida de oportunidades, vid., YZQUIERDO TOLSADA, M., *op.cit.*, pp. 193-195.

Cuestión distinta es que en el ámbito de la IA algunos de esos daños, como los *daños morales puros*, los derivados del *incumplimiento de la normativa de seguridad en materia de IA* y los *causados por la pérdida o corrupción de datos*, cobren un mayor protagonismo que otros y requieran una regulación más estricta para minimizarlos y proteger a los afectados debido a su impacto directo en los derechos fundamentales, la mayor dificultad para repararlos y la expansión de la IA en áreas críticas ¹⁰⁰³. Pensemos en que el uso de estos sistemas de IA es cada vez más frecuente en sectores sensibles como la salud, la justicia o la seguridad, donde los errores pueden tener consecuencias graves al afectar a bienes jurídicos como la dignidad, la intimidad o la salud, que no siempre podrán repararse a través de compensaciones pecuniarias.

Casos de difamaciones llevadas a cabo por sistemas de IA generativa que afectan al honor de las personas, errores en diagnósticos médicos automatizados o fallos de seguridad en vehículos autónomos que provocan daños a la vida e integridad física, o pérdidas de información personal que perjudican la privacidad, no son infrecuentes en la actualidad y todo parece indicar que irán en aumento en la era de la IA.

3. MECANISMOS DE REPARACIÓN DE LOS DAÑOS DERIVADOS DE LA DISCRIMINACIÓN ALGORÍTMICA

Las Directivas de igualdad ordenan a los Estados el establecimiento de sanciones «*efectivas, proporcionadas y disuasorias*» para el caso de que se incum-

¹⁰⁰³ Sobre la pérdida de datos como daño indemnizable, autores como GÓMEZ LIGÜERRE, C.I., «La Propuesta de Directiva sobre responsabilidad ...», *cit.*, y ATIENZA NAVARRO mantienen que los datos en sí mismos no son indemnizables bajo esta Propuesta, sino que lo que se indemnizan son los daños materiales que ocasione dicha pérdida. Como señala el Considerando 20 PDPD, «también debe compensarse la destrucción o corrupción de datos, como los archivos digitales borrados de un disco duro, incluido el coste de recuperar o restaurar los datos [...]», y se aclara que dicha destrucción o corrupción de datos «no da lugar automáticamente a una pérdida material si la víctima puede recuperar los datos sin coste, como por ejemplo cuando existe una copia de seguridad de los datos o los datos pueden descargarse de nuevo, o si un operador económico restaura o restablece los datos temporalmente no disponibles, por ejemplo en un entorno virtual». No obstante, en el caso de que los datos fueran personales, esa pérdida o su manipulación, deterioro o corrupción, constituyen en opinión de RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 657, «tratamientos de datos personales y si se llevan a cabo con infracción de la normativa en este sector con causación de daños, su resarcimiento se podrá fundar en el artículo 82 RGPD. En la determinación de los sujetos responsables, el artículo 82 RGPD limita la legitimación pasiva a los responsables y encargados del tratamiento, y, en consecuencia, en el marco de una tecnología de inteligencia artificial específica deberá valorarse contextualmente si el proveedor o el usuario-operador de esta pueden asumir alguna de estas dos posiciones jurídicas a la hora de responder civilmente con arreglo a este régimen».

plan las obligaciones que imponen¹⁰⁰⁴. Estas disposiciones se presentan como normas de mínimos que permiten a los Estados miembros establecer el elenco de sanciones y reparaciones que consideren oportuno.

Como apunta el TJUE, se ha de determinar en cada caso concreto si las sanciones previstas cumplen dichos requisitos (exigibles incluso cuando no haya una víctima identificable) atendiendo a las circunstancias individuales concurrentes¹⁰⁰⁵. Se entenderá que la sanción es «efectiva» cuando permita compensar a la víctima y castigar al causante; será «proporcionada» cuando su severidad sea acorde a la gravedad de la infracción, es decir, cuando haya un equilibrio entre la sanción y el daño ocasionado; y se entenderá que es «disuasoria» cuando con ella se consigan desalentar posibles conductas discriminatorias previniendo daños futuros¹⁰⁰⁶.

El catálogo de posibles remedios a imponer por los Estados cuando se infringe la normativa es amplio y con ellos se pueden cumplir distintas funciones dependiendo del sector del ordenamiento en que nos encontremos¹⁰⁰⁷. Así, mientras que en el ámbito civil la finalidad será conseguir la reparación

¹⁰⁰⁴ Vid., por ejemplo, el art. 15 de la Directiva 2000/43, por lo que se refiere a la discriminación por origen racial o étnico; el art. 17 de la Directiva 2000/78 por lo que refiere al resto de las discriminaciones en el empleo y la ocupación; el art. 8.2 de la Directiva 2004/113.

¹⁰⁰⁵ En el caso TJUE, C-81/12, *Asociația Accept/Consiliul Național pentru Combaterea Discriminării*, 25 de abril de 2013, el TJUE declaró que correspondía al órgano jurisdiccional nacional valorar si, a tenor de las circunstancias del caso, se satisfacían estos criterios. También dejó claro que cada una de las vías de reparación contempladas en la normativa nacional para los casos de discriminación debía satisfacer, por sí misma, los criterios de efectividad, proporcionalidad y carácter disuasorio.

¹⁰⁰⁶ REY MARTÍNEZ, F., *op.cit.*, p. 128; también puede verse el *Manual de legislación europea contra la discriminación*, cit., p. 283.

¹⁰⁰⁷ Algunas de las sanciones impuestas en los diversos ordenamientos cuando se vulnera el derecho a la igualdad y no discriminación son, entre otras, las siguientes: las multas e indemnizaciones por daños materiales y/o morales, la publicación de la decisión, la prohibición de ejercicio de una profesión o actividad que dependa de la autorización pública, la orden de recolocar al trabajador arbitrariamente despedido, la revisión de los procedimientos de reclutamiento o promoción en la empresa, la formación específica en temas de igualdad a los directivos de las empresas, la suspensión de licencias de actividad o negocio, el restablecimiento de la situación anterior a la discriminación, la prohibición de recibir beneficios públicos o la obligación de establecer planes de igualdad para la entidad, sin perjuicio de las sanciones penales que se prevean. En este sentido, REY MARTÍNEZ, F., *op.cit.*, p. 128. Como explica este autor, a pesar de todas estas posibles sanciones previstas en los ordenamientos nacionales, el análisis del Derecho comparado pone de manifiesto que existen una pluralidad de obstáculos que dificultan la garantía efectiva de este derecho (pueden mencionarse, por ejemplo, la ignorancia sobre los propios derechos o garantías, la complejidad y costes de los procedimientos, la carencia de ayuda legal gratuita, la insuficiente compensación de daños inmateriales o las dificultades en torno a la prueba de la discriminación, todos ellos expuestos en este trabajo). Apunta en la p. 130, que el informe de la ECRI destaca que las autoridades españolas no han informado sobre si las reglas de la responsabilidad civil o administrativa relativas a la indemnización podrían aplicarse en casos de discriminación y tampoco han indicado jurisprudencia en la materia, lo

del daño a la víctima (función reparadora o compensatoria), en el ámbito penal será castigar a los autores de la discriminación (función punitiva). Sin perjuicio de que en ambos casos se persiga, además, prevenir la causación de daños (función preventiva).

Son los Estados lo que habrán de elegir qué medidas consideran más adecuadas para lograr los objetivos marcados por las Directivas. No obstante, como apunta el TJUE en el asunto Arjona-Camacho, dichas medidas tendrán que ser apropiadas para restablecer la igualdad efectiva de oportunidades, garantizar la tutela judicial efectiva y surtir un efecto disuasorio real. Entre las posibles sanciones a imponer se puede incluir la indemnización a la víctima, en cuyo caso, la misma deberá ser *adecuada al daño causado y no meramente simbólica* para garantizar su carácter efectivo y efecto disuasorio, y deberá conseguir la *reparación integral* del perjuicio sufrido¹⁰⁰⁸.

Pues bien, del art. 27 Ley 15/2022 se desprende que, si bien el legislador ha optado por sancionar la discriminación con el pago de una indemnización, éste no es el único mecanismo contemplado para reparar a las víctimas¹⁰⁰⁹. De hecho, en tanto que la misma se refiere exclusivamente a una compensación económica, podría ser insuficiente para lograr la «reparación plena y efectiva» a la que obliga el art. 25.2¹⁰¹⁰. Por ello, la persona física o jurídica que discrimine injustamente quedará obligada *a reparar el daño causado mediante la indemnización correspondiente, pero además, junto con ella y siempre que sea posible, deberá restituir a la víctima a la situación anterior* al incidente discriminatorio.

que hace sospechar que en la práctica no se está aplicando la posibilidad de sancionar por conductas discriminatorias.

¹⁰⁰⁸ Ap. 31 de la STJUE, C-407/14, asunto *María Auxiliadora Arjona Camacho/Securitas Seguridad España, SA*, 17 de diciembre de 2015. Señala el Tribunal que lo que deben introducir en su ordenamiento los Estados Miembros son las medidas que consideren adecuadas para garantizar una indemnización a la víctima *que cubra íntegramente el perjuicio sufrido*.

¹⁰⁰⁹ En la Exposición de Motivos de la Ley se resumen todas las medidas con las que se trata de garantizar este derecho que se reconocen en el Capítulo I del Título II. Esas medidas de protección comprenden: la declaración de nulidad, cese, reparación, prevención, indemnización de daños materiales y morales (en este último caso, en línea con la jurisprudencia más reciente del Tribunal Supremo y Constitucional), así como disposiciones relativas a la tutela judicial y actuación administrativa contra la discriminación, reconociendo en ambos ámbitos respectivamente, una legitimación colectiva a una serie de entidades y organizaciones que tengan entre sus fines la defensa y protección de los derechos humanos.

¹⁰¹⁰ GARCÍA RUBIO, M.P., «Discriminación por razón de sexo ...», *cit.*, p. 162, mantiene que no se debe excluir para este caso «el principio general dimanante del art. 1107 del Código civil, conforme al cual la indemnización debe procurar la reparación integral del daño en el caso del incumplidor doloso, y la reparación de los daños previsibles en el caso del deudor de buena fe [...] la indemnización ha de cubrir, o bien todos los daños sufridos por la víctima a causa de la discriminación intencional, o los que hubieran sido previsibles en el momento de violación de la obligación de no discriminación, cuando no concurre el elemento doloso».

Asimismo, entre las consecuencias jurídicas del acto discriminatorio también se contempla en el art. 26 la *nulidad de pleno derecho* de las disposiciones, actos o cláusulas de los negocios jurídicos que constituyan o causen discriminación.

Por lo que se refiere a la tutela judicial del derecho de igualdad, teniendo en cuenta que las Directivas en materia de no discriminación también obligan a los Estados miembros a establecer *procedimientos judiciales y/o administrativos* que permitan a los particulares hacer valer los derechos reconocidos en ellas¹⁰¹¹, el art. 28 establece que la tutela judicial de este derecho comprenderá *la adopción de todas las medidas necesarias para poner fin a la discriminación* y, en particular, las dirigidas al *cese inmediato* de la discriminación o a la *adopción de medidas cautelares para prevenir violaciones inminentes o ulteriores*, la *indemnización de los daños y perjuicios* causados y el *restablecimiento de la persona perjudicada en el pleno ejercicio de su derecho*.

También el art. 25.1 Ley 15/2022 obliga a la «*articulación de medidas adecuadas para el cese de las situaciones discriminatorias*», derivándose responsabilidades administrativas, así como, en su caso, penales y civiles por los daños y perjuicios que el incumplimiento de dicha obligación pueda ocasionar. Asimismo, el art. 25.3 establece que «ante un incidente discriminatorio, las autoridades encargadas de hacer cumplir esta ley *tomarán las medidas oportunas para garantizar que los hechos no vuelvan a repetirse, especialmente en los casos en los que el agente discriminador sea una administración pública*». Y el art. 31 dispone que la autoridad pública que, con ocasión del ejercicio de sus competencias, tenga conocimiento de la discriminación, deberá, si es competente, incoar el correspondiente procedimiento administrativo en el que se podrán acordar las medidas necesarias para investigar las circunstancias del caso y *adoptar las medidas oportunas y proporcionadas para su eliminación*.

Se observa, pues, que la obligación de reparar el daño causado por la discriminación ha de entenderse en sentido amplio, esto es, como «eliminación de la ilegalidad»¹⁰¹², y puede hacerse efectiva a través de diversas pretensiones y no sólo mediante la compensación económica, posibilidad que en el ámbito de la IA, y más concretamente, de la discriminación algorítmica, no es en ab-

¹⁰¹¹ Vid., art. 9.1 Directiva relativa a la igualdad de trato en el empleo; art 17.1 Directiva sobre igualdad de trato entre hombres y mujeres (refundición); art. 8.1 Directiva sobre igualdad de trato entre hombres y mujeres en el acceso a bienes y servicios; o art. 7.1 Directiva sobre igualdad racial.

¹⁰¹² Para BALLESTER PASTOR, M. A., «Reparación y sanciones en los asuntos de discriminación», Barcelona, 6 de Mayo de 2016, accesible en http://www.era-comm.eu/oldoku/Adiskri/04_Remedies/BALLESTER_116DV37_ES_Paper.pdf, pp. 16 y ss, el concepto «reparación» comprende no solo la compensación económica, sino también la restitución o reinstauración de la víctima en su derecho mediante una obligación de hacer impuesta al ofensor (por ejemplo, mediante la obligación de cesación del acto, la obligatoria readmisión, etc...).

soluto desdeñable teniendo en cuenta que en muchos casos los daños individuales causados por estos sistemas serán poco sustanciales y la indemnización no será el remedio más eficaz para repararlos¹⁰¹³. Por ello, el recurso a la nulidad del acto, decisión o conducta discriminatoria o la posibilidad de exigir en vía judicial o administrativa su cese, pueden tener un valor añadido para el demandante y motivarle para poner en marcha el sistema de responsabilidad civil aun cuando los daños efectivamente sufridos sean de poca entidad¹⁰¹⁴.

Al análisis de estos mecanismos de reparación del daño dedicaremos las siguientes páginas.

3.1. LA INDEMNIZACIÓN POR DAÑOS Y PERJUICIOS CAUSADOS POR LA DISCRIMINACIÓN: ¿LÍMITES INDEMNIZATORIOS?

Quien discrimina en sus relaciones jurídicas ha de reparar el daño que el acto o conducta discriminatoria genere a través de la indemnización. Así lo establece el art. 27 Ley 15/2022 cuando dispone, en primer lugar, que «[l]a persona física o jurídica que cause discriminación por alguno de los motivos previstos en el apartado 1 del artículo 2 de esta ley *reparará el daño causado proporcionando una indemnización* [...]».

Este remedio se contempla igualmente en el art. 72 LOI donde se establece que, *sin perjuicio de otras acciones y derechos contemplados en la legislación civil y mercantil*, la persona que en el ámbito de aplicación del art. 69 LOI (esto es, suministro de bienes o servicios disponibles para el público, ofrecidos fuera del ámbito de la vida privada y familiar) sufra una conducta discriminatoria, tendrá derecho a una *indemnización por los daños y perjuicios sufridos*.

Para GARCÍA RUBIO, se trata de una *acción de responsabilidad contractual específica* y no de una acción de responsabilidad contractual que se solape con la del art. 1.101 CC o con la responsabilidad aquiliana del art. 1.902 CC, aunque podría entrar en concurso con ellas. Además, en su opinión, se trata de una acción que *no presupone como requisito la concurrencia de culpa*, ya que, si así fuera,

¹⁰¹³ Considera con acierto RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 677, que «la reparación de las consecuencias de una discriminación no ha de pasar necesariamente por la utilización del remedio indemnizatorio: otros remedios de conducta pueden ser suficientes para eliminar todas las consecuencias de un acto discriminatorio y restaurar la situación previa a este».

¹⁰¹⁴ Así lo considera RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 671. También plantea este autor la conveniencia de compensar los daños de escasa cuantía a través de la «mera disculpa» del demandado y cita una cuestión prejudicial pendiente ante el TJUE en el ámbito de la protección de datos personales, en la que se plantea si la reparación del daño moral sufrido puede realizarse por medio de una disculpa del responsable del tratamiento de datos ilícito.

se estaría vulnerando el Derecho comunitario teniendo en cuenta la jurisprudencia sentada por el TJUE¹⁰¹⁵. No obstante, esta postura no es pacífica entre los autores, principalmente por dos razones: por un lado, porque en nuestro sistema jurídico el criterio general de imputación de la responsabilidad es la culpa, salvo que el legislador prevea una responsabilidad objetiva o resulte claramente que esa es su intención, o que los daños causados provengan de la creación de un riesgo por parte del obligado. No dándose ninguna de estas circunstancias, hay que entender que la responsabilidad es *subjetiva o por culpa* y procederá por no haber respetado el principio de igualdad de trato en el ámbito de que se trate¹⁰¹⁶; por otro lado, porque en el supuesto especial de responsabilidad civil más próximo al analizado, como es el que deriva de las intromisiones ilegítimas en el derecho al honor, intimidad y propia imagen, la doctrina mayoritaria ha defendido que se trata de un supuesto de responsabilidad civil por culpa¹⁰¹⁷.

Sin perjuicio de lo anterior, hay que tener en cuenta que cuando la discriminación provenga de un sistema de IA público, podrá surgir responsabi-

¹⁰¹⁵ Según observa GARCÍA RUBIO, M.P., «Discriminación por razón de sexo ...», *cit.*, p. 161, la ley no exige ni la intención ni la culpa para considerar que existe discriminación (cf. art. 6 de la Ley Orgánica 3/2007) y si tal requisito subjetivo se exigiese se estaría violando el Derecho comunitario, ya que el TJUE tiene establecido que la responsabilidad derivada del incumplimiento del principio de igualdad de trato deriva de la discriminación, sin que puedan añadirse exigencias ulteriores por el Derecho nacional, como se recoge en *Dekker v. Stichting Vermigenscentrum loor Jong Volwassenen*, asunto C-177/88 (1990) y, más en concreto, no cabe exigir culpa en el autor de la discriminación, como se afirma directamente en *Draehmpaehl v. Urania Immobilienservice OHG*, asunto C 180/95 (1997); en términos muy similares y citando la misma jurisprudencia, AGUILERA RULL, A., «Discriminación directa e indirecta», *InDret* 1/2007, pp. 4-5, mantiene que no se exige en el causante la intención de discriminar y ni siquiera que su conducta sea culpable. De hecho, considera la autora que exigir este requisito atendiendo a la doctrina sentada en el asunto *Draehmpaehl* donde el TJUE consideró que una norma nacional que sometía la indemnización de los daños derivados de una discriminación a la condición de que la conducta hubiera sido culpable, no era conforme a la Directiva 76/207/CEE. En consecuencia, incluso cuando la diferencia de trato pretenda la protección del afectado, podrá considerarse discriminatorias si objetivamente supone una desventaja para la víctima.

¹⁰¹⁶ Opinión que mantiene DÍAZ ALABART, S., «La responsabilidad civil por discriminación ...», *cit.*, apoyándose en numerosa doctrina. Para ITURMENDI MORALES, G., *op.cit.*, p. 14, «los requisitos de la responsabilidad civil por discriminación son los mismos que deben concurrir en el esquema clásico de la responsabilidad civil en general. Es decir, acción u omisión antijurídica de discriminación, *culpabilidad*, resultado dañoso y nexo de causalidad entre la acción u omisión y el daño irrogado al perjudicado».

¹⁰¹⁷ En este sentido, PEÑA LÓPEZ, F., *La culpabilidad en la responsabilidad civil extracontractual*, Granada, 2002, pp. 183-184, citando numerosos autores partidarios de que la LO 1/1982 establece un régimen de responsabilidad subjetivo y también autores cuyos trabajos estudian la responsabilidad objetiva en la legislación española y no incluyen la LO 1/1982 en la enumeración de normas que responden a ese criterio de imputación. En sentido contrario, O'CALLAGHAN MUÑOZ, X., «Compendio de Derecho civil», T.II, vol.2º, 2005, p. 792, considera que se trata de una responsabilidad objetiva, aunque sin razonarlo.

lidad patrimonial para la Administración y la obligación de reparar el daño conforme a los arts. 32 y ss. de la Ley 40/2015, de Régimen Jurídico del Sector Público (LRJSP). Junto con el art. 106.2 CE, dichos preceptos regulan los principios y fundamentos de exigibilidad de la responsabilidad administrativa clásica (esto es, objetivo-patrimonial), recogiendo la Ley 39/2015, las normas relativas al procedimiento para reclamar¹⁰¹⁸. Asimismo, la normativa propia del sector público habrá de conjugarse con el futuro régimen general de responsabilidad por uso de IA que llegue a elaborarse y aprobarse.

En otro orden de cosas, habría que plantearse si los daños derivados de la discriminación serían indemnizables fuera cual fuera su entidad, o si, por el contrario, en ciertos casos las víctimas deberían asumirlos, bien por superar cierto umbral, bien por no alcanzarlo.

Para dar respuesta a esta cuestión debe partirse de que, tanto con carácter general en el ámbito de la responsabilidad por culpa, como concretamente en materia de igualdad y no discriminación, rige el *principio de reparación integral o reparación plena y efectiva* en base al cual *no se pueden establecer límites apriorísticos a la indemnización por daños*. Así lo establecen tanto las propias Directivas europeas como el TJUE en su jurisprudencia¹⁰¹⁹. Dispone expresamente el art. 8.2 *in fine* de la Directiva 2004/113 que «*la fijación de un límite máximo predeterminado no limitará dicha compensación o indemnización*». En consecuencia, se impide a los Estados miembros dar validez a cualquier tipo de límite cuantitativo apriorístico¹⁰²⁰. Esto significa que, si por ejemplo el sistema de IA de un banco niega un préstamo a una clienta por el hecho de ser mujer, y, en consecuencia, no puede celebrar un contrato para el que requería el dinero, el banco

¹⁰¹⁸ En base al mandato constitucional del art. 106.2 según el cual «*los particulares, en los términos establecidos por la ley, tendrán derecho a ser indemnizados por toda lesión que sufran en cualquiera de sus bienes y derechos, salvo en los casos de fuerza mayor; siempre que la lesión sea consecuencia del funcionamiento de los servicios públicos*», y a las concretas previsiones del art. 32 LRJSP, cuyo primer apartado establece que «*Los particulares tendrán derecho a ser indemnizados por las Administraciones Públicas correspondientes, de toda lesión que sufran en cualquiera de sus bienes y derechos, siempre que la lesión sea consecuencia del funcionamiento normal o anormal de los servicios públicos salvo en los casos de fuerza mayor o de daños que el particular tenga el deber jurídico de soportar de acuerdo con la Ley [...]*», la responsabilidad de la Administración se ha configurado como una responsabilidad objetiva. Por tanto, cualquier daño derivado del funcionamiento de los servicios públicos debe ser en principio indemnizado, con independencia de que derive de una actividad normal o anormal de la Administración en la prestación de los servicios públicos, bastando que entre dicha actuación y el daño se dé un nexo causal. Sobre el tema puede verse, ALCOLEA AZCÁRRAGA, C., *op.cit.*, o MARTÍNEZ GUTIÉRREZ, R., «Responsabilidad administrativa por el uso de la Inteligencia Artificial», en GAMERO CASADO/PÉREZ GUERRERO (coords.), *Inteligencia artificial y sector público: retos, límites y medios* Valencia, 2023, pp. 227-253.

¹⁰¹⁹ Vid., la STJUE, C-407/14, asunto *María Auxiliadora Arjona Camacho/Securitas Seguridad España*, SA, 17 de diciembre de 2015.

¹⁰²⁰ Sobre el tema puede verse, BALLESTER PASTOR, M. A., *op.cit.*, pp. 18-19.

debería indemnizar todos los daños derivados de esa pérdida del negocio¹⁰²¹, sin que pudieran imponerse techos o topes máximos a la cantidad que habría de abonar a la mujer si se prueba la discriminación. Con esta norma se trata de garantizar que la víctima reciba una compensación que realmente repare el perjuicio.

También el art. 6.2 de la derogada Directiva 2002/73 y el art.18 de la Directiva 2006/54 establecen como regla general que la indemnización o reparación no puede quedar limitada por un tope máximo fijado *a priori*, salvo un caso muy singular relativo al acceso al empleo¹⁰²². Se hace expresa referencia a que no se cumple adecuadamente con la normativa antidiscriminatoria si los Estados establecen indemnizaciones tasadas para el caso de actuaciones discriminatorias, prohibición coherente con el objetivo de reparación real y efectiva pretendido por dichas normas.

En nuestro Derecho interno, fue el art. 18 de la derogada LIONDAP el que recogió por primera y única vez la prohibición de fijar topes máximos a las indemnizaciones. Dicha prohibición no aparece expresamente ni en la LOI ni en la Ley 15/2022, pero a pesar de esta laguna creemos que una norma legal o contractual que estableciese techos a las indemnizaciones sería nula por contravenir lo dispuesto con carácter imperativo en las Directivas comunitarias en perjuicio de sus destinatarios.

Además, respecto a los techos indemnizatorios en casos de daños ocasionados por la IA, debe tenerse en cuenta que la nueva Directiva de responsabilidad por productos defectuosos establece en su Considerando 56 que, en ese ámbito, no se podrá excluir o limitar la responsabilidad de un operador económico mediante disposiciones contractuales ni del Derecho nacional, ya que, de ser así, podría menoscabarse la protección de las personas físicas. En consecuencia, el art. 15 obliga a los Estados miembros a garantizar que la responsabilidad del operador respecto al perjudicado, no quede limitada o excluida por tales disposiciones. En base a esta previsión normativa, los legisladores nacionales no podrán establecer techos indemnizatorios para limitar los importes de los daños derivados de un defecto de producto, aunque podrían preverlos para otros regímenes de responsabilidad civil que afectaran a sistemas de IA.

Por otro lado, habría que plantearse si sería necesario que los daños fueran significativos o superaran cierto umbral para ser resarcibles, sobre todo

¹⁰²¹ GARCÍA RUBIO, M.P., «Discriminación por razón de sexo ...», *cit.*, p. 164.

¹⁰²² El último inciso de dichos preceptos dispone que, excepcionalmente, la indemnización tasada sí que será válida en aquellos casos en que el empresario pueda probar que el único perjuicio sufrido por el demandante como resultado de la discriminación es la negativa a tomar en consideración su solicitud de trabajo.

teniendo en cuenta que en el ámbito de la IA el impacto que los resultados de los sistemas pueden tener en la esfera individual puede ser de escasa entidad, por ejemplo, un pequeño sobrecoste en la compra de bienes con precios personalizados discriminatorios¹⁰²³.

En España, a diferencia de otros ordenamientos jurídicos donde no es posible reclamar una indemnización por daños insignificantes o que no superen cierto umbral, contamos con un sistema abierto de reparación del daño o de cláusula general¹⁰²⁴. Por tanto, en principio no habría inconveniente en que dichos daños fueran indemnizables. Además, parece lógico pensar que si no se exige probar el daño moral porque el mismo se presume, con menor razón se exigirá acreditar que el mismo es de cierta entidad o que supera cierto umbral.

En el ámbito de la protección de datos personales, el TJUE ha mantenido que una norma nacional que exigiera un cierto umbral de relevancia para la reparación de los daños inmateriales sería contraria al art. 82 RGPD, ya que éste no lo exige y se limita a enumerar los tipos de daños resarcibles sin establecer umbrales¹⁰²⁵. Por lo que se refiere a las vulneraciones al derecho a la igualdad y no discriminación, consideramos que sería extrapolable esta misma argumentación al no haber ninguna norma que establezca dicho umbral y al operar el principio de reparación integral del daño. Esta conclusión no es baladí teniendo en cuenta que no será infrecuente que los daños que ocasionen la discriminación algorítmica en la esfera patrimonial y/o moral de un sujeto sean mínimos.

3.1.1. La importancia de las acciones colectivas en la lucha contra la discriminación algorítmica

En ocasiones, la escasa entidad de los daños sufridos por las víctimas individuales de discriminación algorítmica en comparación con el elevado coste que podría alcanzar la presentación de una demanda de responsabilidad, podría desincentivar el ejercicio de acciones. E incluso cuando los daños fueran de cierta gravedad, la complejidad y opacidad de los sistemas de IA podría encarecer tanto estos procedimientos al ser preciso contar con peritos especializados en detectar y probar sus fallos, que podría augurarse que en no pocos

¹⁰²³ RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 630.

¹⁰²⁴ Ello no obsta a que algunos regímenes jurídicos concretos, como el de productos defectuosos, establezcan *topes mínimos* para indemnizar.

¹⁰²⁵ Vid., la STJUE de 4 de mayo de 2023, en el asunto C-300/21, *UI v. Österreichische Post AG* (ECLI:EU:C:2023:370).

casos los costes de la reclamación judicial podrían superar los propios daños sufridos y los beneficios que se espera obtener con las demandas¹⁰²⁶.

Ahora bien, siendo cierto que en ocasiones el daño individual podría ser menor, no lo será el global o agregado teniendo en cuenta la escala y velocidad con que pueden adoptarse decisiones discriminatorias a través de estos sistemas. En palabras de RUBÍ PUIG, «si se toma en consideración la acumulación de pérdidas individuales, la externalidad total producida puede ser significativa»¹⁰²⁷.

Precisamente por ello, se ha mantenido con acierto por algunos autores la conveniencia de hacer frente a los daños derivados de la IA en general, y de la discriminación algorítmica en particular, mediante *acciones colectivas* que ayuden a las víctimas a simplificar los trámites burocráticos y reducir los costes económicos derivados de la defensa de sus derechos (honorarios de abogados y procuradores, peritajes, posibles costas, etc.), contribuyendo así a la reparación de los daños simbólicos¹⁰²⁸.

Llama la atención que ni el RIA ni las Directivas propuestas en materia de responsabilidad prevean la posibilidad de ejercitar acciones colectivas y delegar dichas acciones en una entidad, organización o asociación, a diferencia de lo previsto por ejemplo en el RGPD o en las propias Directivas de igualdad, que exigen a los Estados miembros que garanticen que las asociaciones, organizaciones o personas jurídicas con interés legítimo puedan iniciar, en

¹⁰²⁶ RUBÍ PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 665, explica con claridad esta idea en relación a los sistemas de IA de personalización de precios: «Si el sistema de inteligencia artificial utilizado en la personalización infringe de algún modo la normativa destinada a hacer estos sistemas más fiables, seguros, transparentes y no discriminatorios, habrá motivos para hacer responder al proveedor u operador del sistema de los daños causados. Pero estos daños en la esfera patrimonial individual serán mínimos. El afectado que quizá haya pagado 10 euros más por un producto o que a consecuencia de la oferta personalizada mostrada finalmente no llegó a contratar tendrá pocos incentivos a reclamar. El esfuerzo que le supondría sería las más de las veces muy superior al beneficio esperado de su reclamación».

¹⁰²⁷ *Ibidem*.

¹⁰²⁸ GERARDS, J./XENIDIS, R., *op.cit.*, p. 76 y p. 145. También RUBÍ PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 671. El hecho de que en Europa no se hayan interpuesto prácticamente dichas acciones es uno de los impedimentos más importantes para la protección del derecho a la igualdad y no discriminación. Así lo corrobora SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación...», *cit.*, p. 18/31, citando numerosos estudios sobre el tema. Entre ellos, es muy interesante el trabajo de NAGY, C. I., *Collective Actions in Europe: A Comparative, Economic and Transsystemic Analysis*, Cham, Springer, 2019, donde se ponen de relieve varias ideas importantes. Por ejemplo, se destaca la eficiencia de dichas acciones cuando las reclamaciones individuales son excesivamente costosas para los particulares y se proponen mecanismos para su implementación (pp. 9-20), o se ponen de manifiesto las principales objeciones contra este mecanismo, entre las que se mencionan la falta de tradición en el Derecho continental, los problemas para identificar a los miembros del grupo afectado, o el temor a demandas masivas (vid., pp. 23-42).

nombre del afectado o en su apoyo, y con su autorización, cualquier procedimiento judicial o administrativo para exigir que se cumplan las obligaciones que en ellas se imponen¹⁰²⁹.

Por lo que se refiere al Derecho interno, esta obligación se contempla en el art. 31 Ley 62/2003, en el art. 12 LOI y en el más reciente art. 29 Ley 15/2022. Este último precepto, además de reconocer legitimación individual a los propios afectados para la defensa de su derecho a la igualdad y no discriminación, *legitima a los partidos políticos, sindicatos, asociaciones profesionales de trabajadores autónomos, organizaciones de consumidores y usuarios, y asociaciones u organizaciones legalmente constituidas que tengan entre sus fines la defensa y promoción de los derechos humanos*, para defender, en los términos establecidos por las leyes procesales, los derechos e intereses de las personas afiliadas o asociadas o usuarias de sus servicios en procesos judiciales civiles, contencioso-administrativos y sociales, siempre que cuenten con su autorización expresa. Además, exige a las asociaciones y organizaciones que defienden y promueven los derechos humanos, la acreditación de una serie de requisitos relacionados, por un lado, con el momento de su válida constitución y el ejercicio de su actividad, y por otro lado, con su ámbito de actuación.

La DF 2^a de la Ley 15/2022 modificó también el art. 11 bis de la LEC sobre la legitimación para la defensa del derecho a la igualdad de trato y no discriminación, que ahora regula la cuestión en los términos siguientes:

«1. Para la defensa del derecho a la igualdad de trato y no discriminación, además de las personas afectadas y siempre con su autorización, estarán también legitimados la Autoridad Independiente para la Igualdad de Trato y la No Discriminación, así como, en relación con las personas afiliadas o asociadas a los mismos, los partidos políticos, los sindicatos, las asociaciones profesionales de trabajadores autónomos, las organizaciones de personas consumidoras y usuarias y las asociaciones y organizaciones legalmente constituidas que tengan entre sus fines la defensa y promoción de los derechos humanos, de acuerdo con lo establecido en la Ley integral para la igualdad de trato y la no discriminación. 2. Cuando las personas afectadas sean una pluralidad indeterminada o de difícil determinación, la legitimación para instar acciones judiciales en defensa de derechos o intereses difusos

¹⁰²⁹ Vid., por ejemplo, art. 7 Directiva 2000/43/CE, art. 9 Directiva 2000/78/CE o art. 17 Directiva 2006/54/CE. Asimismo, señala COTINO HUESO, L., «Discriminación, sesgos e igualdad de la inteligencia artificial en el sector público», en GAMERO CASADO/PÉREZ GUERRERO (coords.), *Inteligencia artificial y sector público: retos, límites y medios* Valencia, 2023, p. 337, que podría llegar a ser aplicable en su caso la Directiva (UE) 2020/1828 del Parlamento Europeo y del Consejo de 25 de noviembre de 2020 relativa a las acciones de representación para la protección de los intereses colectivos de los consumidores, y por la que se deroga la Directiva 2009/22/CE («Directiva 2020/1828»).

corresponderá a la Autoridad Independiente para la Igualdad de Trato y la No Discriminación, a los partidos políticos, los sindicatos y las asociaciones profesionales de trabajadores autónomos más representativos, así como a las organizaciones de personas consumidoras y usuarias de ámbito estatal, a las organizaciones, de ámbito estatal o del ámbito territorial en el que se produce la situación de discriminación que tengan entre sus fines la defensa y promoción de los derechos humanos, de acuerdo con lo establecido en la Ley integral para la igualdad de trato y la no discriminación, sin perjuicio en todo caso de la legitimación individual de aquellas personas afectadas que estuviesen determinadas».

Del análisis de ambos preceptos (art. 29 Ley 15/2022 y art. 11 bis LEC) podemos extraer las siguientes ideas:

La primera, que a las personas legitimadas por el art. 29 Ley 15/2022 se añade en el art. 11 bis LEC la *Autoridad Independiente para la Igualdad de Trato y la No Discriminación*, organismo independiente y unipersonal, basado fundamentalmente en la *auctoritas* de su titular, cuya principal misión es proteger frente a la discriminación y promover el cumplimiento del derecho antidiscriminatorio y entre cuyas funciones está precisamente la de ejercitar acciones judiciales en defensa de los derechos derivados de la igualdad de trato y la no discriminación (*vid.*, art. 40.1.d). Se le atribuye legitimación extraordinaria por sustitución que le permite actuar en el proceso en nombre e interés propio, pero en ejercicio de un derecho ajeno¹⁰³⁰.

La segunda, que además de regular la legitimación procesal en los supuestos en que los afectados están determinados o son fácilmente determinables (esto es, cuando se pretenden instar acciones judiciales en defensa de *derechos o intereses colectivos*), el art. 11 bis LEC también contempla la legitimación para la defensa de este derecho en los casos en que los perjudicados son indeterminados (esto es, cuando se trata de defender *derechos o intereses difusos*), previsión especialmente interesante en el ámbito de la IA. En el primer caso, además, la legitimación que se otorga a dichas entidades se hace concretamente respecto de sus afiliados o asociados por cuyo nombre y cuenta se actúa.

En tercer lugar, que se exigen una serie de requisitos para reconocer legitimación a dichos entes. Para empezar, ambos preceptos exigen la autorización de la persona afectada. Ahora bien, mientras el art. 29 Ley 15/2022 dispone que la misma ha de ser «expresa», el art. 11 bis LEC no contempla esta exigencia. Además, el art. 29.2 añade otros dos requisitos que deben acreditar las asociaciones y organizaciones de defensa y promoción de los derechos humanos: pri-

¹⁰³⁰ CHOCRÓN GIRÁLDEZ, A.M., *op.cit.*, p. 99. Cuando se trate de defender intereses colectivos necesitará la autorización de las personas afectadas para ser parte en el proceso por discriminación.

mero, haberse constituido legalmente al menos dos años antes de la iniciación del proceso judicial y venir ejerciendo de modo activo las actividades necesarias para alcanzar los fines previstos en sus estatutos, salvo que ejerciten las acciones administrativas o judiciales en defensa de los miembros que la integran; y segundo, que desarrollen su actividad en el ámbito estatal o, en su caso, en un ámbito territorial que resulte afectado por la posible situación de discriminación¹⁰³¹.

Por otro lado, dispone también el art. 31 Ley 15/2022 que «a los efectos de lo establecido en el artículo 4.2 de la Ley 39/2015, de 1 de octubre, del Procedimiento Administrativo Común de las Administraciones Públicas, los sindicatos, las asociaciones profesionales de trabajadores autónomos, las organizaciones de personas consumidoras y usuarias y las asociaciones y organizaciones legalmente constituidas que tengan entre sus fines la defensa y promoción de los derechos humanos y cumplan los requisitos fijados en el artículo 29 de la presente ley, *podrán tener la consideración de interesado en los procedimientos administrativos en los que la Administración tenga que pronunciarse en relación con una situación de discriminación prevista en esta ley*, siempre que cuenten con la autorización de la persona o personas afectadas. No será necesaria esta autorización cuando las personas afectadas sean una pluralidad indeterminada o de difícil determinación, sin perjuicio de que quienes se consideren afectados puedan también participar en el procedimiento».

A la vista de tales previsiones normativas, queda claro que la actuación individual en defensa del derecho a la igualdad y no discriminación no empece a que dicha protección pueda encomendarse también a una persona distinta del afectado. Esta legitimación colectiva presenta múltiples ventajas tanto para los particulares como para el propio sistema judicial, ya que, reduce costes derivados de los pleitos y permite que personas con pocos recursos puedan acceder a la justicia, simplifica la gestión de los casos permitiendo una resolución más rápida y eficiente de los conflictos y contribuye a la aplicación uniforme de la ley. En definitiva, contribuye al respeto de principios como el de economía procesal y el de homogeneidad de las resoluciones judiciales.

Asimismo, estas acciones pueden tener un impacto sistémico en la erradicación de prácticas discriminatorias. Por un lado, porque a raíz de su interposición, las empresas u organizaciones pueden quedar obligadas a revisar y modificar sus políticas o prácticas discriminatorias. Junto con ello, porque la

¹⁰³¹ En los supuestos en que se trate de defender intereses difusos, el art.11 bis establece unas reglas particulares de legitimación en función de la entidad de que se trate. Así, en el caso de las organizaciones de personas consumidoras y usuarias, la legitimación activa se reduce a aquéllas que sean de ámbito estatal (art. 11 bis 2 LEC); y en el caso de sindicatos y las asociaciones profesionales de trabajadores autónomos la legitimación activa se reconoce a los tengan la condición de más representativos (art. 11 bis 2 LEC).

entidad de las posibles sanciones y la publicidad negativa asociada a una acción colectiva pueden tener un considerable efecto disuasorio contra la discriminación. Por último, porque las sentencias que resuelvan estos casos podrán establecer importantes precedentes que beneficiarán no sólo a los demandantes sino también a otros individuos en situaciones similares, lo que redundará en el desarrollo y consolidación de una jurisprudencia sólida en materia de igualdad y no discriminación.

Además, mediante la interposición de estas demandas colectivas se puede contribuir al empoderamiento de ciertos grupos vulnerables, ya que se visibilizan problemas sistémicos de discriminación y se fortalece la cohesión social entre los propios colectivos afectados. Dicho empoderamiento puede ser crucial para promover cambios sociales y jurídicos.

Tales ventajas han llevado a la doctrina a defender la conveniencia de que en las futuras regulaciones en materia de IA se incluyan normas específicas que armonicen mínimamente el régimen procesal de las acciones colectivas, evitando que su disponibilidad y características queden sujetas a la decisión particular de cada Estado miembro¹⁰³².

No obstante, ha de tenerse en cuenta que pocos días antes de cerrar este trabajo, se publicaba en el BOE (fecha 14 de marzo), el texto del nuevo Proyecto de Ley de acciones colectivas para la protección y defensa de los derechos e intereses de los consumidores y usuarios, con el que se pretende trasponer, con el habitual retraso, la Directiva (UE) 2020/1828, del Parlamento Europeo y del Consejo de 25 de noviembre de 2020, relativa a las acciones de representación para la protección de los intereses colectivos de los consumidores. Dicha Directiva pretende garantizar que todos los Estados miembros tengan, al menos, un mecanismo procesal efectivo para que las organizaciones de consumidores puedan iniciar acciones contra empresas que infrinjan el Derecho de la Unión en materia de consumo.

En esta línea, con el Proyecto de ley se pretende mejorar la regulación actual en materia de tutela judicial colectiva de los derechos de los consumidores y usuarios, mediante la creación de un procedimiento especial, dentro del Libro IV de la Ley de Enjuiciamiento Civil, que aborde de forma sistemática todas las reglas procesales que requiere el adecuado tratamiento normativo de las acciones colectivas. Además, se propone un ámbito de aplicación amplio, que dé cobertura al ejercicio de acciones frente a cualquier tipo de infracción en que se hayan visto perjudicados los derechos e intereses colectivos de los consumidores y usuarios

¹⁰³² En este sentido, RUBI PUIG, A., «Inteligencia artificial y daños indemnizables», *cit.*, p. 671, considera muy conveniente que, igual que se hizo en el art. 80 RGPD, se prevean normas sobre el ejercicio de acciones colectivas en este sector.

(vid., Exposición de Motivos). Como novedad, se regula el ejercicio de acciones tanto para detener prácticas ilícitas (acciones colectivas de cesación) como para obtener indemnizaciones (acciones colectivas resarcitorias).

3.1.2. Compatibilidad con la acción de responsabilidad del art. 82 RGPD

Como se apuntó *supra*, el art. 27 de la Ley 15/2022 permite exigir una indemnización al responsable de la discriminación cuando ésta queda acreditada, con independencia de su causa u origen. Ahora bien, cuando el resultado discriminatorio de un sistema de IA trae causa de un tratamiento ilícito de datos personales, ha de tenerse en cuenta que a través de una misma operación se estarían infringiendo simultáneamente dos derechos fundamentales (el de igualdad y no discriminación, y el de protección de datos de carácter personal) y entraría en escena también la normativa de protección de datos de carácter personal. De hecho, en los considerandos del RGPD se pone de relieve que los tratamientos de datos pueden conllevar riesgos de gravedad y probabilidad variables para los derechos y libertades de las personas, pudiendo ocasionar *problemas de discriminación*, entre otros que se particularizan¹⁰³³. Esto puede ocurrir, como sigue diciendo el Considerando 75, «[...] *en los casos en los que los datos personales tratados revelen el origen étnico o racial, las opiniones políticas, la religión o creencias filosóficas, la militancia en sindicatos y el tratamiento de datos genéticos, datos relativos a la salud o datos sobre la vida sexual, o las condenas e infracciones penales o medidas de seguridad conexas*».

De forma más específica, las *Directrices sobre la evaluación de impacto relativa a la protección de datos (EIPD) y para determinar si el tratamiento «entraña probablemente un alto riesgo» a efectos del Reglamento (UE) 2016/679* –Directrices WP24823–,

¹⁰³³ Dispone el Considerando 75 que «[l]os riesgos para los derechos y libertades de las personas físicas, de gravedad y probabilidad variables, pueden deberse al tratamiento de datos que pudieran provocar daños y perjuicios físicos, materiales o inmateriales, en particular en los casos en los que el tratamiento pueda dar lugar a problemas de discriminación, usurpación de identidad o fraude, pérdidas financieras, daño para la reputación, pérdida de confidencialidad de datos sujetos al secreto profesional, reversión no autorizada de la seudonimización o cualquier otro perjuicio económico o social significativo; en los casos en los que se prive a los interesados de sus derechos y libertades o se les impida ejercer el control sobre sus datos personales; en los casos en los que los datos personales tratados revelen el origen étnico o racial, las opiniones políticas, la religión o creencias filosóficas, la militancia en sindicatos y el tratamiento de datos genéticos, datos relativos a la salud o datos sobre la vida sexual, o las condenas e infracciones penales o medidas de seguridad conexas; en los casos en los que se evalúen aspectos personales, en particular el análisis o la predicción de aspectos referidos al rendimiento en el trabajo, situación económica, salud, preferencias o intereses personales, fiabilidad o comportamiento, situación o movimientos, con el fin de crear o utilizar perfiles personales; en los casos en los que se traten datos personales de personas vulnerables, en particular niños; o en los casos en los que el tratamiento implique una gran cantidad de datos personales y afecte a un gran número de interesados».

adoptadas por el GT29 en 4 de abril de 20217, interpretan que la protección se ha de extender a otros derechos fundamentales, señalando explícitamente la libertad de expresión, la libertad de pensamiento, la libertad de circulación, *la prohibición de discriminación* y la libertad de conciencia y de religión¹⁰³⁴.

Y respecto de las brechas de datos personales y seguridad en los tratamientos, es decir, de los tratamientos no autorizados o accidentales sobre los datos, dispone el Considerando 85 que, si no se adoptan a tiempo medidas adecuadas, estas violaciones de la seguridad de los datos personales pueden entrañar daños y perjuicios físicos, materiales o inmateriales para las personas físicas, como pérdida de control sobre sus datos personales o restricción de sus derechos, *discriminación*, usurpación de identidad, pérdidas financieras, reversión no autorizada de la seudonimización, daño para la reputación, pérdida de confidencialidad de datos sujetos al secreto profesional, o cualquier otro perjuicio económico o social significativo.

Como pone de manifiesto la AEPD en su guía sobre *Gestión del Riesgo y evaluación de impacto en tratamientos de datos personales*, de junio de 2021, también los fallos y errores técnicos derivados de la implementación automatizada de tratamientos de datos personales que se pudieran derivar de, por ejemplo, sistemas de decisión automática, sistemas de ayuda a la decisión, tratamiento biométrico, etc., pueden ocasionar problemas específicos de confidencialidad, integridad, disponibilidad, trazabilidad o autenticación, *pero también otros más específicos relacionados con la calidad de los datos o la discriminación*, y estima necesario que se estudien de forma particular a la hora de analizar el riesgo para los derechos y libertades¹⁰³⁵.

Asimismo, dicha Agencia destaca como factores de riesgo no explícitos en la normativa de protección de datos, incluyéndolos entre los efectos colaterales, *la posible inferencia de categorías especiales de datos a partir de la información acumulada del usuario, la discriminación en la oferta de opciones, productos o servicios a causa del perfilado del usuario, los sesgos en la toma de decisiones o la discriminación algorítmica*.

Queda patente, de todo lo anterior, la estrecha interconexión que existe entre el tratamiento de datos personales y la posible discriminación que puede sufrir una persona.

¹⁰³⁴ Vid., p. 7. A tenor del art. 35 RGPD, cuando sea probable que un tipo de tratamiento entrañe un alto riesgo para los derechos y libertades de las personas físicas, el responsable del tratamiento realizará, antes del tratamiento, una evaluación del impacto de las operaciones de tratamiento en la protección de datos personales. Según indica el GT29, la referencia que hace el art. 35 RGPD a «los derechos y libertades» de los interesados atañe principalmente a los derechos a la protección de datos y a la intimidad, pero también puede implicar otros derechos fundamentales como la libertad de expresión, la libertad de pensamiento, la libertad de circulación, la prohibición de discriminación, el derecho a la libertad y la libertad de conciencia y de religión.

¹⁰³⁵ P. 43.

Siendo esto así, hay que plantearse la posible compatibilidad de las acciones de responsabilidad previstas en las distintas normativas aplicables (en concreto, la del art. 82 RGPD y la del art. 27 Ley 15/2022) para conseguir la reparación de los daños derivados de las decisiones discriminatorias de los sistemas de IA que además se basan en datos personales.

Cada una de estas normativas tiene un ámbito de aplicación distinto y persigue fines distintos. Y aunque ambas regulan acciones para reclamar responsabilidad civil por daños, lo hacen con un enfoque diferente dado que el bien jurídico que pretenden proteger también lo es. Así, mientras la Ley 15/2022 se centra en la protección de la igualdad y en la discriminación como ilícito principal a combatir, siendo la afectación de los datos personales un medio para causar el daño y no el elemento central de la infracción, el RGPD se fundamenta en la protección de los datos personales y el daño ha de surgir del tratamiento ilícito de los mismos, sin que necesariamente deba existir una discriminación. No obstante, ambas normativas pueden solaparse en los casos en que el uso indebido de datos personales tenga un efecto discriminatorio, en cuyo caso el mismo hecho podría dar lugar a reclamaciones bajo regímenes diferentes.

El ejercicio de cada una de estas acciones queda subordinado a la vulneración de derechos distintos y a la concurrencia de presupuestos o requisitos propios para su interposición. Así, el art. 27 Ley 15/2022 obliga al causante de la discriminación a indemnizar a la víctima por los daños y perjuicios sufridos como consecuencia de cualquier conducta o acto discriminatorio (incluidos los que deriven del tratamiento de datos personales), y el art. 82 RGPD reconoce a los titulares de datos personales una acción indemnizatoria que pueden ejercer contra los responsables y encargados del tratamiento para resarcirse de daños y perjuicios materiales e inmateriales sufridos a consecuencia de las operaciones que estos realicen contraviniendo dicha normativa.

Los requisitos de la acción de responsabilidad civil por discriminación coinciden con los presupuestos clásicos de la cláusula general de responsabilidad. Es decir, ha de concurrir la acción u omisión antijurídica de discriminación, la culpabilidad, el daño (con las particularidades en este ámbito respecto de la presunción de existencia de daños morales) y el nexo de causalidad entre la acción u omisión y el daño irrogado al perjudicado¹⁰³⁶. Y para que prospere una acción de responsabilidad civil en virtud del art. 82 RGPD el afectado deberá acreditar la infracción de la normativa sobre protección de datos, los daños y perjuicios materiales o inmateriales sufridos, la condición de responsable y/o encargado del tratamiento de datos del sujeto causante de

¹⁰³⁶ ITURMENDI MORALES, G., *op.cit.*, p. 14.

los daños, y la existencia de una relación de causalidad entre el incumplimiento de la normativa y el perjuicio ocasionado.

Dándose los presupuestos descritos, no vemos inconveniente en que las acciones de responsabilidad derivadas del art. 27 de la Ley 15/2022 y del art. 82 del RGPD pudieran coexistir, siempre y cuando no se diera una duplicidad indemnizatoria que supusiera un enriquecimiento injusto para el perjudicado.

Doctrina y jurisprudencia se han manifestado a favor de la compatibilidad de acciones cuando un mismo hecho vulnera el derecho a la protección de datos y el derecho al honor, la intimidad o la propia imagen¹⁰³⁷. Similares consideraciones podrían hacerse, a nuestro juicio, en relación al derecho de igualdad y no discriminación. Téngase en cuenta, además, que no solo no existe una exclusión normativa que impida la interposición simultánea de ambas acciones, sino que el Considerando 146 RGPD recoge la posibilidad de que el perjudicado reclame en virtud del art. 82 RGPD una indemnización por los daños y perjuicios sufridos como consecuencia de un tratamiento que infrinja el Reglamento, pero «*sin perjuicio de cualquier reclamación por daños y perjuicios derivada de la vulneración de otras normas del Derecho de la Unión o de los Estados miembros*».

No obstante, creemos que esta acumulación de acciones (y su éxito) exigiría acreditar un daño diferenciado bajo cada marco normativo, lo que no siempre será sencillo. Es decir, cuando la decisión discriminatoria a la que llega un sistema de IA es consecuencia de un tratamiento ilícito de datos o proviene de la infracción de la normativa al efecto (por ejemplo,

¹⁰³⁷ Vid., MORENO MARTÍNEZ, J.A., «El impacto del Reglamento General de Protección de Datos en el régimen de responsabilidad civil (art. 82 RGPD): Su posible desarrollo por el Derecho interno y problemática de coexistencia con otros mecanismos protectores», en ATAZ LÓPEZ, J. y COBACHO GÓMEZ, J.A. (Coords.), *Cuestiones clásicas y actuales del Derecho de daños: estudios en homenaje al profesor Dr. Roca Guillamón*, Tomo 3, 2021, pp. 515-565. Afirma este último autor que la compatibilidad indemnizatoria ha sido aceptada reiteradamente por nuestros tribunales, ya que, aunque se sustentan en fundamentos jurídicos distintos «la similitud de la pretensión indemnizatoria perseguida justifica que el alcance del daño pueda verse proyectado a ambos fundamentos, como resultado de ser vulnerados»; RUBÍ PUIG, A., «Problemas de coordinación y compatibilidad entre la acción indemnizatoria del artículo 82 del Reglamento General de Protección de datos y otras acciones en derecho español», *Derecho Privado y Constitución*, n. 34, 2019, pp. 223 y ss.

La acumulación de ambas acciones por un mismo hecho dañoso se ha reconocido, por ejemplo, en los supuestos de inclusión indebida en registros de morosos o en los supuestos de vulneración del derecho al olvido. Respecto a los primeros, vid., la STS 826/2022, de 24 de noviembre, en la que se declara que cuando mediante la vulneración de la normativa de protección de datos resulte afectado el derecho al honor de una persona será indemnizable la afectación a la dignidad en su aspecto interno y externo. En mismo sentido, la STS 267/2023, de 20 de febrero (ECLI:ES:TS:2023:989) o la STS 130/2020, de 27 de febrero (RJ 2020/613).

porque el uso ilegítimo de categorías especiales de datos ha conllevado que el sistema adopte decisiones basadas en el origen étnico o racial de las personas, que perjudican a los miembros de ese grupo), será complejo diferenciar el daño ocasionado por la discriminación de la persona (a la que, por ejemplo, se le deniega un préstamo personal por motivos raciales), del daño que pueda sufrir por la infracción del art. 9 RGPD. Para ser resarcido por ambas vías, habría que probar que el hecho de haber utilizado categorías especiales de datos sin base legitimadora para ello (a pesar de la prohibición del art. 9 RGPD), ocasionó daños propios y distintos a los que ocasionó la discriminación. En el ejemplo expuesto, el daño patrimonial o moral sufrido por la persona provendría principalmente de la denegación del crédito por motivos discriminatorios, denegación a la que el sistema de IA llega porque se introducen como variables a tener en cuenta categorías de datos protegidas como la raza o etnia de los solicitantes. A nuestro modo de ver, el daño patrimonial que podría ocasionar la denegación del préstamo (por ejemplo, la imposibilidad de emprender un negocio o la pérdida de una oportunidad de inversión), así como el daño moral derivado de la lesión de la dignidad de la persona discriminada sería, a su vez, el daño que derivaría también de la infracción de la normativa sobre protección de datos. Sería difícil distinguir otros daños derivados del tratamiento ilegítimo de tales datos adicionales a los ocasionados por la propia discriminación¹⁰³⁸. Solo cuando esto quedara acreditado por el demandante, podría tener éxito la acción de responsabilidad del art. 82 RGPD.

Ha de tenerse en cuenta aquí que, como ha establecido el TJUE en su sentencia de 4 de mayo de 2023, asunto C-300-21 *Österreichische Post AG*, la mera infracción del RGPD no basta para que surja responsabilidad civil del infractor, sino que será necesario probar que la misma ocasionó daños al interesado. En otras palabras, el TJUE reconoce que la indemnización regulada en el art. 82 RGPD queda supeditada a que la persona haya sufrido daños y

¹⁰³⁸ Cabe mencionar aquí el caso resuelto por el TS en su sentencia de 5 de abril de 2016 (RJ 2016/1006) en relación con el derecho al olvido, donde se alega vulneración tanto del derecho a la protección de datos como de los derechos de la personalidad protegidos por la LO 1 /1982. Respecto a los daños morales que se acreditaron en el procedimiento, el Supremo hace las siguientes apreciaciones: por un lado, que los daños morales se reconocen en función de las distintas vulneraciones de derechos que se estiman, en concreto, la vulneración a los derechos al honor, intimidad e imagen y el derecho a la protección de datos personales; por otro lado, y esto es lo que ahora nos interesa, mantiene que a pesar de que concurre la vulneración de varios derechos, la conducta ilícita es una y el daño moral causado también es único. En palabras del Tribunal, «se trata de un concurso ideal con relación a una sola conducta y a un único resultado lesivo que debe ser indemnizado con criterios estimativos». En función de esto, no se fijan indemnizaciones distintas o cuantías diferentes por cada uno de los derechos vulnerados (tal como pretendía el actor), sino que se fija una cantidad global por el daño moral.

perjuicios como consecuencia de la infracción del RGPD, pero no cuando a pesar de dicha trasgresión normativa el interesado no haya sufrido daño alguno. Ello sin perjuicio de las correspondientes reclamaciones que pudieran presentarse ante las autoridades de control *ex arts. 77 y 78 RGPD*, o las sanciones o multas que pudieran corresponder por vulnerar dicha normativa en virtud de los arts. 83 y 84 RGPD¹⁰³⁹.

En definitiva, cuando quede acreditado que a través de una misma operación de un sistema de IA se han vulnerado de forma simultánea ambos derechos fundamentales, aunque la víctima podría recurrir a la acción más favorable para sus intereses, en la práctica se podrían plantear ambas conjuntamente en vía judicial. Dichas acciones serían compatibles ya que protegen derechos distintos. No obstante, acudir al art. 27 tendría la ventaja de que el demandante podría beneficiarse de la presunción de existencia de daños morales una vez acreditada la discriminación¹⁰⁴⁰. Según una opinión doctrinal, esto podría implicar la aplicación indirecta de una presunción *iure et de iure* contenida en otra norma para la prueba de los daños y perjuicios derivados del tratamiento de datos personales que podría alterar el régimen de carga de la prueba propio del RGPD, lo que podría interpretarse como contrario al RGPD por el TJUE¹⁰⁴¹. No obstante, otros consideran que cuando existe discriminación a través del tratamiento

¹⁰³⁹ Como señala MORENO MARTÍNEZ, J.A., *op.cit.*, p. 553, basándose en la jurisprudencia del TS, al margen de la posibilidad de interponer una acción de responsabilidad civil ante los tribunales, el afectado por el tratamiento de datos podrá iniciar también un procedimiento administrativo sancionador ante la autoridad de control en virtud del art. 77 RGPD a fin de que se adopten medidas para su cese tras la declaración de la infracción, sin perjuicio además de las posibles sanciones que puedan corresponder. No obstante, concluye que no toda infracción en el tratamiento por parte del responsable o encargado implica necesariamente la causación de daños al titular de los datos, sino que se exige la prueba de la relación de causalidad y, por ende, su incidencia perjudicial.

¹⁰⁴⁰ Respecto a la presunción contenida en el art. 9.3 LO 1/1982, considera RUBÍ PUIG, A., «Problemas de coordinación y compatibilidad...», *cit.*, pp. 223-224, que teniendo en cuenta dicha presunción de existencia de daños morales no será extraño que «en el marco de una acción contra un responsable o un encargado del tratamiento, los abogados de los afectados por una infracción de la normativa sobre protección de datos intenten también recurrir a los remedios previstos en la LO 1/1982 para defender la posición de su cliente y acumular ambas acciones». Además, para dicho autor, la existencia de esta presunción unida a otros factores como la experiencia acumulada por los tribunales españoles en la aplicación de la normativa recogida en la Ley 1/1982 y las incertezas sobre la coordinación entre las acciones del RGPD y los procedimientos seguidos ante la AEPD u otras autoridades de control, podrían menoscabar la utilidad práctica de la acción del art. 82 RGPD. Según su opinión, «la posibilidad de acumular ambas acciones podría llevar a que los tribunales españoles, por inercia, siguieran dando prioridad a los mecanismos previstos en la LO 1/1982 y relegaran la aplicación del art. 82 del RGPD a casos muy específicos en los que el derecho de daños no es un medio adecuado de compensación».

¹⁰⁴¹ Así lo entiende RUBÍ PUIG, A., «Problemas de coordinación y compatibilidad...», *cit.*, p. 224, respecto a la análoga presunción contenida en el art. 9.3 LO 1/1982. En sentido contrario, MORENO MARTÍNEZ, J.A., *op.cit.*, p. 564, no cree que esta aplicación contradiga los principios del

de datos resultaría plenamente aplicable la presunción de daño moral del art. 27.1 de la Ley 15/2022, ya que no solo se estaría vulnerando el derecho a la protección de datos sino también el derecho a no ser discriminado, razón por la que también entienden que si el tratamiento de datos vulnerara el derecho al honor, la intimidad o la propia imagen, igualmente resultaría de aplicación la presunción de daño del art. 9.3 LO 1/1982¹⁰⁴².

Respecto a la legitimación pasiva, mientras el RGPD establece que será el responsable o el encargado del tratamiento el que deba indemnizar cualesquiera daños y perjuicios que pueda sufrir una persona a consecuencia del incumplimiento de los deberes que dicha normativa les impone, el obligado a reparar el daño ocasionado por la discriminación será, en virtud del art. 27 Ley 15/2022, la persona física o jurídica que cause la discriminación por alguno de los motivos previstos en el apartado 1 del artículo 2 de dicha ley. En los casos a los que hemos venido haciendo referencia en este trabajo, normalmente coincidirán en la misma persona la condición de responsable del tratamiento y la de causante de la discriminación, teniendo en cuenta que el empresario (público o privado) que adopta una decisión discriminatoria mediante o con apoyo en un sistema de IA que decide implementar en su empresa, sería tanto el autor de la discriminación, como el responsable del tratamiento de datos en la medida en que es quien decide los fines y medios de dicho tratamiento (art. 4.7 RGPD) y el que decide utilizar el sistema de IA en su ámbito de actuación, por lo que podría resultar civilmente responsable en base a ambos marcos normativos.

Por último, a pesar de sus diferencias, ambas acciones de responsabilidad también tienen algunos rasgos comunes. Por lo que se refiere a los daños indemnizables, ambas regulaciones cubrirían tanto los daños materiales como inmateriales sufridos por la persona. Además, se basan en el *principio de reparación integral o reparación plena y efectiva* del daño, tal como se reconoce en el Considerando 146 RGPD y en el art. 25 de la Ley 15/2022.

RGPD al encontrarse dentro de las especificidades que pueden existir en los Estados miembros sobre las acciones judiciales, tal como reconoce el propio Reglamento.

¹⁰⁴² CASADESÚS RIPOLL, P., *op.cit.*, p. 123. No obstante, como reconoce la autora, esta opinión no es pacífica. Defendiendo la aplicación de la presunción del art. 9.3 LO 1/1982 a los daños derivados de un tratamiento de datos personales, también GRIMALT SERVERA, P., *La responsabilidad civil en el tratamiento automatizado de datos personales*. Granada, 1999, p. 140. También MORENO MARTÍNEZ, J.A., *op.cit.*, p. 556, para quien, la compatibilidad de estas acciones conlleva claros efectos beneficiosos como son, por un lado, *la aplicación de la presunción de existencia de daños del art. 9 LO 1/1982*, y por otro lado, la posibilidad de que puedan adoptarse simultáneamente junto al resarcimiento solicitado, medidas tendentes al cese de la intromisión ilegítima y al restablecimiento del afectado al estado anterior a su vulneración. Precisamente ambos beneficios aconsejan la conciliación de las acciones y su enjuiciamiento conjunto.

Asimismo, sin perjuicio de la posibilidad de reclamar una indemnización por daños, habría que tener en cuenta la posibilidad de ejercitar otras acciones o exigir responsabilidades en virtud del RIA, a los distintos operadores de la tecnología por el uso de sistemas de IA de alto riesgo que lleguen a decisiones sesgadas o discriminatorias.

3.2. NULIDAD DE PLENO DERECHO

La discriminación constituye un motivo de nulidad de pleno derecho de las disposiciones, actos o cláusulas de los negocios jurídico-privados que la ocasionen. Así lo dispone el art. 26 de la Ley 15/2022 integral de igualdad, en virtud del cual serán «nulos de pleno derecho las *disposiciones, actos o cláusulas de los negocios jurídicos* que constituyan o causen discriminación por razón de alguno de los motivos previstos en el apartado primero del artículo 2 de esta ley».

La redacción de este precepto es parcialmente coincidente con la recogida en el art. 10 LOI, titulado «Consecuencias jurídicas de las conductas discriminatorias», según el cual «[l]os *actos y las cláusulas de los negocios jurídicos* que constituyan o causen discriminación por razón de sexo *se considerarán nulos y sin efecto*, y darán lugar a *responsabilidad* a través de un sistema de reparaciones o indemnizaciones que sean *reales, efectivas y proporcionadas al perjuicio sufrido*, así como, en su caso, a través de un *sistema eficaz y disuasorio de sanciones* que prevenga la realización de conductas discriminatorias». Dicho precepto fue muy criticado porque adolecía de ciertas deficiencias técnicas que dificultaban su interpretación:

En primer lugar, porque su título no se correspondía con su contenido, ya que mientras aquél aludía con carácter general a las consecuencias jurídicas de cualquier «conducta» discriminatoria, éste hacía referencia exclusivamente a «actos y cláusulas» de los negocios jurídicos, lo que dejaba fuera un buen número de casos que no podían incardinarse en dicha categoría jurídica. A fin de soslayar esta incoherencia, la doctrina consideró que debía hacerse una interpretación extensiva del término «acto» que diera cabida a cualquier actuación discriminatoria¹⁰⁴³. Esta conclusión se apoyaba también en una interpretación sistemática de la norma, ya que la LOI, a lo largo de su articulado, aludía también a «actividades», «transacciones» o «conductas» que iban más allá de la mera actuación negocial, aunque también incluyera ésta (*vid.*, arts. 69 y 72.1).

¹⁰⁴³ En este sentido, por todos, GARCÍA RUBIO, M.P., «Discriminación por razón de sexo ...», *cit.*, p. 136 y NAVAS NAVARRO, S., «Negativa a contratar y prohibición de discriminar...», *cit.*, p. 1624.

Por su parte, el art. 26 Ley 15/2022 incluye junto con la expresión «actos y cláusulas» de los negocios jurídicos, también las «disposiciones» que constituyan o causen discriminación. Dicho término también se emplea en algunas Directivas comunitarias que contemplan la nulidad de las «*disposiciones* contrarias al principio de igualdad que figuren en los *contratos* individuales o en los *convenios colectivos*, en los *reglamentos internos* de las empresas, en las *normas que rijan las asociaciones* con o sin ánimo de lucro, así como en los *estatutos* de las profesiones independientes y de las organizaciones sindicales y empresariales»¹⁰⁴⁴. No se incluyen, sin embargo, las decisiones unilaterales del empresario que pueden ser uno de los focos más relevantes de tratamientos discriminatorios.

Estas lagunas e incoherencias normativas han llevado a la doctrina a interpretar, en aras de garantizar en mayor medida el derecho a la igualdad y no discriminación, que *cualquier conducta discriminatoria* entendida ésta en sentido amplio como comprensiva tanto de prácticas sistemáticas, como de conductas espontáneas, decisiones, disposiciones, actos o cláusulas contractuales y cualquier otra actuación o comportamiento que ocasione discriminación injusta, podría declararse nula por vulnerar dicho derecho fundamental. Extrapolando esto al ámbito de los sistemas de IA, podría declararse nula la salida generada por el sistema en forma de predicción, recomendación, decisión, etc., siempre que fuera discriminatoria.

En segundo lugar, el art. 10 LOI establece con un alcance muy general, criticable y criticado, toda una amalgama de remedios de diversa índole ante una conducta discriminatoria dirigidos a garantizar la indemnidad de la víctima en todos los ámbitos del ordenamiento¹⁰⁴⁵. Así, junto a consecuencias típicamente civiles como la nulidad y la responsabilidad por daños, se contemplan sanciones penales y administrativas dirigidas a prevenir la discriminación, esto es, con eficacia *ex ante* en lugar de *ex post*.

Parece que el legislador de 2022, consciente de que no era adecuado aunar en el mismo precepto una pluralidad de medidas de distinta naturaleza y fines, ha mejorado la estructura de la regulación contemplando separadamente en el art. 26 la nulidad de pleno derecho, y en los artículos siguientes otros remedios como la indemnización, la restitución de la víctima a la situación anterior a la discriminación o la acción de cesación.

Por lo que se refiere concretamente a la *nulidad de pleno derecho*, ha sido criticada por ser una medida excesivamente genérica que impide atender a

¹⁰⁴⁴ Vid., art. 14 de la Directiva 43/2000, por lo que refiere a la discriminación por origen racial o étnico; art. 15 de la Directiva 2000/78 por lo que refiere al resto de las discriminaciones en el empleo y la ocupación; o art. 5 de la Directiva 76/207, sobre igualdad de trato entre hombres y mujeres.

¹⁰⁴⁵ Criticando la norma, por todos, GARCÍA RUBIO, M.P., «Discriminación por razón de sexo...», *cit.*

las peculiaridades de muchos supuestos y satisfacer los intereses de las víctimas¹⁰⁴⁶. Teniendo en cuenta que esta sanción implica la ineficacia de la actuación discriminatoria con la consiguiente restitución de las partes a la situación anterior al acto declarado nulo, la doctrina pone en tela de juicio que sea la medida más beneficiosa para quien sufre discriminación. Por ejemplo, para la mujer embarazada a la que el Banco ha concedido un préstamo en condiciones sustancialmente peores que las concedidas a un hombre o a una mujer no embarazada, podría ser más beneficioso que dicha entidad quedara obligada a concertar un nuevo préstamo en condiciones no discriminatorias, en lugar de que el contrato se declarara nulo¹⁰⁴⁷; y lo mismo podría decirse en los casos de negativa a contratar por motivos discriminatorios, donde los efectos de la nulidad absoluta tampoco solucionarían el problema del perjudicado ya que sólo implicarían la obligación de reponer a las partes al momento de la elección de la contraparte del contrato, pero no la obligación del causante de la discriminación de contratar a la persona discriminada¹⁰⁴⁸.

Se ha discutido si el particular discriminado podría imponer al autor de la discriminación la celebración del contrato, su celebración en otras condiciones o incluso la modificación del contrato ya celebrado. Es decir, la doctrina se plantea si entre los remedios a disposición de la persona discriminada para conseguir el cese de la discriminación, estaría el *derecho a celebrar el contrato* o a *hacerlo en condiciones más beneficiosas*, ya que su admisión mermaría el principio de autonomía de la voluntad, en virtud del cual, los contratantes no sólo son libres para celebrar (o no) un contrato, sino también para elegir a su contraparte y las condiciones en las que se quiere contratar. Por su parte, el *derecho a modificar unilateralmente un contrato ya perfeccionado* vulneraría el principio de obligatoriedad recogido en los arts. 1091 y 1258 Cc.

El art. 69 LOI, tras imponer en su primer inciso la obligación de los suministradores de bienes o servicios disponibles para el público, ofrecidos fuera del ámbito de la vida privada y familiar, de cumplir en sus actividades y transacciones con el principio de igualdad de trato entre mujeres y hombres y de no discriminación por razón de sexo, establece en su segundo apartado que dicha obligación «no afecta a la libertad de contratación, incluida la libertad

¹⁰⁴⁶ ORMAZABAL SÁNCHEZ, G., *Discriminación y carga de la prueba en el proceso civil*, Madrid, 2011, p. 46 y ss., considera justificadas tales críticas.

¹⁰⁴⁷ GARCÍA RUBIO, M.P., «Discriminación por razón de sexo...», *cit.*, p. 158.

¹⁰⁴⁸ Mantiene NAVAS NAVARRO, S., «Negativa a contratar y prohibición de discriminar...», *cit.*, p. 1625, que si fuese ésta la pretensión de la víctima debería fundamentarse mediante la exigencia de responsabilidad (contractual o extracontractual dependiendo de cómo se considere) por el daño causado. En cualquiera de los dos casos, la víctima podría pretender la reparación *in natura* a fin de que se impusiera a la parte discriminadora la obligación de contratar con ella. Sobre cómo se articularía esta medida en uno y otro caso, vid., pp. 1626-7.

de la persona de elegir a la otra parte contratante, *siempre y cuando dicha elección no venga determinada por su sexo*»¹⁰⁴⁹. Una interpretación *a sensu contrario* de la norma llevaría a concluir que cuando dicha elección se basara en el sexo del contratante, sí quedaría afectada su libertad de contratación. Es decir, la libertad de elegir a la contraparte quedaría limitada cuando la misma viniera determinada exclusivamente en base al sexo de la persona.

Así parece entenderlo también NAVAS NAVARRO para quien, en tales casos, no existiría libertad de elección y el autor de la discriminación se vería obligado a contratar con la contraparte si ésta le exigiera la celebración del contrato, siempre y cuando dicha restricción de la libertad respetara el *principio de proporcionalidad* al que debe adecuarse cualquier limitación de derechos fundamentales, es decir, siempre que no hubiera un medio menos gravoso de solucionar el conflicto entre esos bienes jurídicos¹⁰⁵⁰. A la misma conclusión podría llegarse cuando la negativa a contratar se basase en la elaboración de perfiles o toma de decisiones discriminatorias realizados por sistemas de IA. En los casos en que tales sistemas recomendaran o decidieran directamente no ofrecer un bien o servicio a un sujeto y ese resultado automatizado fuera asumido por la contraparte humana, el afectado lo podría impugnar y podría exigir la celebración del contrato, sin perjuicio de la indemnización por daños¹⁰⁵¹.

No obstante, para GARCÍA RUBIO el art. 69.2 no resuelve el problema analizado, sino que, muy al contrario, guarda silencio sobre la tensión entre la libertad contractual y la prohibición de discriminación. Tampoco otras normas nacionales, como las laborales, que establecen la nulidad de la discriminación en el acceso al empleo precisando que la misma dará lugar a las responsabilidades pertinentes, imponen expresamente al empresario la obligación de contratar al trabajador discriminado. No obstante, esto no impediría fundamentar dicha medida por la vía de la responsabilidad civil¹⁰⁵². La respuesta pasaría, en opinión de la autora, por distinguir varios tipos de situaciones:

¹⁰⁴⁹ Este precepto incorpora la disposición contenida en el art. 3.2 de la Directiva 2004/113/CE, de 13 de diciembre, por la que se aplica el principio de igualdad de trato entre hombres y mujeres al acceso a bienes y servicios y su suministro, según el cual: «La presente Directiva no afectará a la libertad de la persona a la hora de elegir a la otra parte contratante, siempre y cuando dicha elección no se base en el sexo de la persona contratante».

¹⁰⁵⁰ Para NAVAS NAVARRO, S., «Negativa a contratar y prohibición de discriminar...», *cit.*, p. 1622, esta norma recoge un límite a la aplicación de los derechos fundamentales entre particulares, límite que podrá imponerse siempre y cuando sea proporcionado. Y citando a DIEZ PIZACO señala que cualquier restricción a los derechos fundamentales «debe adecuarse al *principio de proporcionalidad*, de suerte que la injerencia en la cláusula general sólo procede cuando no se puede alcanzar el fin que se persigue por un *medio menos gravoso*».

¹⁰⁵¹ NAVAS NAVARRO, S., «La perspectiva de género ...», *cit.*, pp. 16 y 17/23.

¹⁰⁵² GARCÍA RUBIO, M.P., «Discriminación por razón de sexo ...», *cit.*, p. 153; también NAVAS NAVARRO, S., «Negativa a contratar y prohibición de discriminar...», *cit.*, p. 1632.

por un lado, aquellas en las que *el contrato ya no es posible o no resulta de utilidad o interés para el sujeto discriminado*; y por otro lado, aquellas en las que *el contrato es todavía posible sin perjudicar a terceros de buena fe*, debiendo limitarse sólo a estos últimos casos la posibilidad de imponer al autor de la discriminación la obligación de contratar con quien no quiere. Además, puesto que admitir sin cortapisas este remedio perjudicaría gravemente la libertad de contratación, la autora defiende su aplicación sólo en ciertos supuestos.

Entre las situaciones en las que no podría imponerse la obligación de contratar estarían, en primer lugar, aquéllas en que el contrato se hubiese celebrado ya con un tercero de buena fe protegido y esto excluyera la posibilidad de contratar con la persona discriminada. Esta situación podría darse, por ejemplo, cuando una persona hubiese arrendado su vivienda a un hombre en detrimento de una mujer, basando su elección exclusivamente en el sexo de los candidatos, y el arrendador ya no tuviera más viviendas disponibles para alquilar. Junto con estos supuestos, estarían aquéllos en que la celebración del contrato en un momento posterior al inicialmente propuesto, carecería de interés o utilidad para la persona discriminada. Siguiendo con el ejemplo anterior, cuando la mujer discriminada ya hubiera arrendado otra vivienda en mejores condiciones que las ofertadas por el discriminador, o cuando ya no le interesara esa vivienda por haber cambiado su lugar de residencia. En estos casos, la vía más idónea para reparar a la víctima sería la indemnización por daños y perjuicios.

Al margen de estos supuestos, podrían darse otros en los que la reparación de la víctima sí podría incluir la obligación de contratar con ella si ésta fuera su voluntad. Sería el caso de que la prestación de bienes o servicios se hiciera en régimen de monopolio, ya que entonces la negativa a contratar implicaría la imposibilidad de acceder a esa prestación al no existir otra alternativa en el mercado¹⁰⁵³; y algo similar podría ocurrir cuando circunstancias extraordinarias y excepcionales impidieran el correcto funcionamiento del mercado, provocando una situación de necesidad justificativa de la imposición de la obligación de contratar para evitar perjuicios al sujeto discriminado¹⁰⁵⁴. En ambos casos, la reparación a través de la obligación de contratar sería una sanción «proporcionada» al daño sufrido, puesto que no habría otra medida alternativa igualmen-

¹⁰⁵³ NAVAS NAVARRO, S., «Negativa a contratar y prohibición de discriminar...», *cit.*, p. 1631, igual que GARCÍA RUBIO, considera que la libertad contractual debe disminuir o desaparecer en la misma medida que disminuye o desaparece la libre competencia. De compatibilidad entre la *libertad de contratación* y la existencia de una *obligación de contratación en casos de monopolio o de bienes socialmente relevantes*, habla AGUILERA RULL, A., «Prohibición de discriminación y libertad de contratación», *InDret* 1/2009, p. 26, para quien también sería compatible dicha libertad con la posible revisión del contenido del contrato en casos de desequilibrio negocial.

¹⁰⁵⁴ GARCÍA RUBIO, M.P., «Discriminación por razón de sexo ...», *cit.*, p. 158.

te eficaz y menos gravosa, ya que, si bien con la indemnización se conseguiría reparar el daño moral derivado de la discriminación, no ocurriría lo mismo respecto del daño material derivado de la imposibilidad de acceder al bien o servicio concreto, sobre todo si es esencial¹⁰⁵⁵. En definitiva, solo en dichos casos sería proporcionado imponer la obligación de contratar, sin perjuicio de la indemnización por daños morales y/o materiales que pudiera corresponder.

Distinto, aunque próximo, sería el caso de que existiera alternativa en el mercado pero fuera de peor condición que la que se deniega injustamente. Para NAVAS NAVARRO, cuya opinión compartimos, puesto que aquí sería posible celebrar un «negocio de sustitución», aun en peores condiciones, no cabría imponer al discriminador la obligación de contratar sino sólo la reparación mediante una indemnización que debería comprender la diferencia pecuniaria satisfecha al haber realizado el negocio en peores condiciones, además de la indemnización del daño moral¹⁰⁵⁶.

Por otro lado, en los casos en que la discriminación no se hubiese materializado en la negativa a contratar sino en *la contratación en condiciones más perjudiciales para la víctima*, la consecuencia sería la recogida en el art. 26 Ley 15/2022, esto es, la nulidad de pleno derecho. Sería el caso en que el propietario del ejemplo anterior arrendase la vivienda a una mujer a un precio muy superior al contenido en la oferta publicada en Internet para el público en general, por el mero hecho de ser mujer.

Idéntica sanción procedería cuando la persona hubiera sido discriminada en base a una decisión arrojada por un sistema inteligente. La víctima podría solicitar la nulidad de los negocios jurídicos celebrados o de alguna de sus cláusulas (en concreto, las discriminatorias), como por ejemplo podría ocurrir cuando se le hubieran exigido mayores garantías para la concesión de un préstamo por ser mujer y tener un trabajo a tiempo parcial.

Se ha puesto en duda la conveniencia de esta sanción civil para las víctimas de la discriminación. Pensemos en el ejemplo del arrendamiento anterior. A

¹⁰⁵⁵ En este sentido, NAVAS NAVARRO, S., «Negativa a contratar y prohibición de discriminar...», *cit.*, pp. 1627 y ss.

¹⁰⁵⁶ Entiende GARCÍA RUBIO, M.P., «Discriminación por razón de sexo...», *cit.*, p. 158, que en aquellos casos en los que quien sufre la discriminación puede lograr idénticos o similares bienes o servicios en el mercado, sería preferible el remedio indemnizatorio a la imposición de la celebración del contrato en contra de la libertad de contratación. En el mismo sentido, NAVAS NAVARRO, S., «Negativa a contratar y prohibición de discriminar...», *cit.*, p. 1632, considera que imponer la obligación de contratar cuando el sujeto discriminado puede disponer de otras alternativas razonables en el mercado («negocio de sustitución»), no es una forma de reparación que sea necesaria, adecuada y proporcionada al daño sufrido, puesto que existe otra medida alternativa que resulta también eficaz y menos gravosa por cuanto con ella no se restringe la libertad contractual, como es la indemnización por daños y perjuicios.

la mujer que celebra un contrato obligándose a pagar una renta muy superior a la ofertada por el propietario al público en general, quizá no le interese su nulidad (que en este caso sería total al afectar la cláusula ilícita a un elemento esencial como el precio, objeto del propio contrato). Lo que realmente le podría ser útil sería que la renta se redujera hasta equipararse a la publicitada en la oferta. Por tanto, sería más efectivo que pudiera instar la modificación del contrato y exigir las mismas condiciones que le hubieran correspondido en ausencia de discriminación¹⁰⁵⁷. Con ello se podría sortear el indeseado efecto que provocaría la nulidad absoluta del contrato, la cual, no lo olvidemos, también podría instar la contraparte que ocasionó la discriminación.

Esta solución la prevé el art. 72.2 LOI para los contratos de seguro o de servicios financieros afines, limitación que suscita dudas sobre la posibilidad de generalizarla a otros supuestos de discriminación, en tanto en cuanto supone un atentado a la eficacia obligatoria de los contratos y no está prevista en el Cc. Tal como dispone dicho precepto, el incumplimiento de la prohibición de establecer diferencias en las primas y prestaciones de las personas aseguradas por razón del sexo, otorgará al contratante perjudicado *el derecho a reclamar la asimilación de sus primas y prestaciones a las del sexo más beneficiado, manteniéndose en los restantes extremos la validez y eficacia del contrato*¹⁰⁵⁸. Para GARCÍA RUBIO se trata de una modificación del contrato de origen legal que conlleva la sustitución de las cláusulas nulas (por discriminatorias *in peius*), por las más favorables aplicadas al otro género¹⁰⁵⁹. Para JIMÉNEZ HORWITZ lo que hace el precepto es optar por la nulidad parcial del contrato mediante la eliminación de las cláusulas discriminatorias ilícitas y su integración con las condiciones de trato igualitario cuando lo solicite el perjudicado, de forma que acabe teniendo acceso a servicios o beneficios negados injustamente¹⁰⁶⁰. A nuestro modo de ver, lo que permite el precepto es que, en un caso en que en principio procedería la nulidad absoluta por infracción de normas imperativas o prohibitivas, se derive un efecto distinto pudiendo optar la víctima de la discriminación por otra opción más beneficiosa (arts. 6.3 y 1255 Cc).

¹⁰⁵⁷ GARCÍA RUBIO, M.P., «Discriminación por razón de sexo...», *cit.*, p. 159, cita doctrina alemana que se decanta por esta posibilidad.

¹⁰⁵⁸ Téngase también en cuenta que la disposición final 13.1 de la Ley 11/2013, de 26 de julio, suprimió la excepción que recogía el párrafo segundo del art.71.1, según la cual, reglamentariamente se podían fijar supuestos en los que fuera admisible determinar diferencias proporcionadas de las primas y prestaciones de las personas consideradas individualmente, cuando el sexo constituyera un factor determinante de la evaluación del riesgo a partir de datos actuariales y estadísticos pertinentes y fiables.

¹⁰⁵⁹ GARCÍA RUBIO, M.P., «La igualdad de trato entre hombres y mujeres y su repercusión en el derecho de contratos...», *cit.*, p. 12/21.

¹⁰⁶⁰ JIMÉNEZ HORWITZ, M., «La protección contra la discriminación en las relaciones entre particulares: la evolución desde la responsabilidad extracontractual hasta los remedios por incumplimiento», *ADC*, tomo LXVII, 2014, fasc. II, p. 503.

Ahora bien, fuera de los concretos supuestos previstos bajo el ámbito de aplicación del art. 72 LOI, ¿podría optar la víctima de la discriminación por esta medida? La reciente Ley 15/2022 integral de igualdad no contempla esta posibilidad, al menos expresamente. No obstante, ¿podría entenderse implícitamente incluida en el art. 28 que permite a los tribunales la adopción de *cuantas medidas sean necesarias* para poner fin a la discriminación de que se trate?; ¿caería dentro de aquéllas dirigidas a conseguir el cese de la discriminación?; ¿podría considerarse que sería una vía para conseguir, como también establece el dicho precepto, el restablecimiento de la persona perjudicada en el pleno ejercicio de su derecho? A nuestro juicio, una interpretación sistemática y finalista de la norma, sólo podría conducir a la respuesta afirmativa de todas estas cuestiones, ya que con esta medida se pondría fin a la discriminación y se conseguiría restablecer al perjudicado en el ejercicio de su derecho a la igualdad de trato, medida que el precepto distingue de la indemnización por daños en sentido estricto.

3.3. ACCIÓN DE CESACIÓN DE LA CONDUCTA DISCRIMINATORIA

Ante una situación de discriminación efectivamente producida, la víctima puede exigir también su cese judicialmente. Esa acción de cesación se contempla en el art. 28 Ley 15/2022 y presenta cierta semejanza con lo dispuesto en el art. 9.2 de la LO 1/1982. A tenor de aquél, «[l]a tutela judicial frente a las vulneraciones del derecho a la igualdad de trato y no discriminación comprenderá, en los términos establecidos por las leyes procesales, la adopción de todas las medidas necesarias para poner fin a la discriminación de que se trate y, en particular, las dirigidas al *cese inmediato de la discriminación, pudiendo acordar la adopción de medidas cautelares dirigidas a la prevención de violaciones inminentes o ulteriores* [...]».

Esta acción tiene por finalidad solicitar al juez que ordene al demandado poner fin a la conducta discriminatoria. Puesto que el precepto transcrito hace referencia al «cese inmediato» de la discriminación, está haciendo referencia a una conducta en curso respecto de la que se reclama su cese. No obstante, además de detener la discriminación actual, esta acción puede incluir medidas para prevenir que se repita en el futuro. Téngase en cuenta que la norma se refiere «a la prevención de violaciones *inminentes o ulteriores*» a cuyo fin plantea la adopción de medidas cautelares. Así, durante el proceso, el demandante también podría solicitar el cese de inmediato de la conducta discriminatoria como medida cautelar mientras se resuelve el fondo del asunto. Con ello podría evitarse la agravación de los daños ocasionados o la causación de otros nuevos¹⁰⁶¹.

¹⁰⁶¹ CHOCRÓN GIRÁLDEZ, A.M., *op.cit.*, p. 94.

El cese de la discriminación podría pasar también por la prohibición de seguir utilizando el sistema de IA sesgado. En este sentido, señala COTINO HUESO que cuando se detecta que un sistema de IA público es discriminatorio deberán adoptarse medidas de mitigación, modulación, corrección o eliminación de las causas, que podrían llegar hasta el bloqueo o el cese del uso del sistema¹⁰⁶². Y explica RUBÍ PUIG que en el ámbito de la IA uno de los debates más incipientes tiene que ver con el denominado remedio a la «destrucción algorítmica», es decir, la pretensión dirigida a modificar o suprimir datos empleados por el algoritmo o de alguno de sus elementos.

Recordemos que, como se explicó en su momento, el RIA impone a los proveedores de sistemas de alto riesgo que detecten que no cumplen con el Reglamento, la obligación de adoptar de forma inmediata las medidas correctoras necesarias para que sean conformes o para retirarlos del mercado, desactivarlos o recuperarlos, según proceda. Además, informarán de ello a los distribuidores y, en su caso, a los responsables del despliegue, al representante autorizado y a los importadores (vid., arts. 16.1. j y 20 RIA). Además, «investigarán inmediatamente, en su caso, las causas» (art. 21) e informarán a los organismos y autoridades competentes (art. 22). La autoridad de vigilancia del mercado que aprecie un riesgo para la igualdad del sistema de alto riesgo tendrá que evaluar el sistema y «exigirá sin demora indebida al agente económico correspondiente que adopte todas las medidas correctoras adecuadas», en su caso, «todas las medidas provisionales adecuadas para prohibir o restringir la comercialización del sistema de IA en su mercado nacional, retirar el producto de ese mercado o recuperarlo». También habrá de informar a las autoridades competentes, incluso advertir a la Comisión y otros Estados si el riesgo va más allá de España (art. 65 RIA)¹⁰⁶³.

Por último, en el contexto de la protección de datos también los responsables y encargados del tratamiento estarían obligados a adoptar medidas en caso de detectar deficiencias del sistema, así como a cumplir las órdenes de las autoridades de control que podrían consistir, por ejemplo, en la limitación temporal o definitiva del tratamiento, incluida su prohibición, tal como dispone la normativa al efecto (art. 58.2.f RGPD).

¹⁰⁶² Para dicho autor, tales obligaciones son claras desde la normativa de igualdad, de procedimiento administrativo, de protección de datos y del RIA. En COTINO HUESO, L., «Discriminación, sesgos e igualdad de la inteligencia artificial en el sector público», *cit.*, p. 330.

¹⁰⁶³ COTINO HUESO, L., «Discriminación, sesgos e igualdad de la inteligencia artificial en el sector público», *cit.*, p. 332.

Bibliografía

- ABRIL STOFFELS, R., «El reconocimiento judicial en la discriminación múltiple en el ámbito europeo», *Revista de Derecho Comunitario Europeo*, n° 44, 2013, pp. 309 y ss.
- AGENCIA DE LOS DERECHOS FUNDAMENTALES DE LA UNIÓN EUROPEA (FRA) y CONSEJO DE EUROPA, *Manual de legislación europea contra la discriminación (edición 2018)*, Luxemburgo, 2019.
- AGENCIA ESPAÑOLA DE PROTECCIÓN DE DATOS (AEPD), *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción*, 2020.
- Guía sobre *Gestión del Riesgo y evaluación de impacto en tratamientos de datos personales*, junio, 2021.
- AGUILERA RULL, A., «Discriminación directa e indirecta», *InDret* 1/2007, pp. 1-17.
- «Prohibición de discriminación y libertad de contratación», *InDret* 1/2009, pp. 1-30.
- «Daño moral por discriminación», en Gómez Pomar, F./Marín García, I. (coords.), *El daño moral y su cuantificación*, Madrid, 2023, pp. 559-586.
- ALAMEDA CASTILLO, M.T., «Reclutamiento tecnológico. Sobre algoritmos y acceso al empleo», *Temas laborales: Revista andaluza de trabajo y bienestar social*, N° 159, 2021, pp. 11-52.
- ALCOLEA AZCÁRRAGA, C., «La responsabilidad patrimonial de la Administración y el uso de algoritmos», *Revista General de Derecho Administrativo*, n° 59, enero 2022.
- ÁLVAREZ CUESTA, H. *El impacto de la IA en el trabajo: desafíos y propuestas*, Aranzadi, Navarra, 2020.
- «Inteligencia artificial: derecho de la UE y derecho comparado. La propuesta de una ley sobre IA», en P. Rivas Vallejo (dir.), *Discriminación algorítmica en el ámbito laboral*, Navarra 2022, pp. 379-414.
- ÁLVAREZ OLALLA, M.P., «Propuesta de Reglamento en materia de responsabilidad civil por el uso de inteligencia artificial, del Parlamento Europeo, de 20 de octubre de 2020», *CESCO de Derecho de Consumo*, N°. 38, 2021, pp. 1-10.
- *Manual de Derecho de Daños*, Navarra, 2021.

- ANGUITA RÍOS, R., «Inteligencia artificial y Derecho civil: líneas de pensamiento en materia de daños», *Revista Crítica de Derecho Inmobiliario*, N.º 781, 2020, pp. 2541-2581.
- ARCOS VARGAS, M., «La no discriminación en el Derecho derivado de la Unión Europea» en J.M. MORALES ORTEGA (dir.), *Realidad social y discriminación. Estudios sobre diversidad e inclusión laboral*, Murcia, 2022, pp. 17-41.
- ARGELICH COMELLES, C., «Deberes de transparencia del Reglamento 2019/1150 (P2B Regulation) para prevenir la discriminación algorítmica del consumidor en los sistemas de prelación de ofertas», *Cuadernos de derecho transnacional*, vol. 15, n. 1, 2023, pp. 129-135.
- ARIAS DOMÍNGUEZ, A., *La cuantificación de la indemnización por daño moral por trasgresión de derechos fundamentales en los despidos nulos*, Colección de Derecho del Trabajo y Seguridad Social (Agencia Estatal Boletín Oficial Del Estado), n.º 15, Madrid, 2023.
- ARROYO AMAYUELAS, E., «El scoring de Schufa», *Indret*, n. 3, 2024, pp. 134-160.
- ASQUERINO LAMPARERO, M.J «Algoritmos y discriminación», en AA.VV. *Realidad social y discriminación. Estudios sobre diversidad e inclusión laboral*, Murcia, Laborum, 2022.
- «Análítica de personas y discriminación algorítmica en procesos de selección y contratación», *Labos*, Vol. 5, Número extraordinario sobre Normativa europea sobre inteligencia artificial, pp. 99-130.
- ATAZ LÓPEZ, J., «Daños causados por las cosas: una nueva visión a raíz de la robótica y de la inteligencia artificial», *Càtedra Jean Monnet Working Paper*; 4/2020, pp.1-57.
- ATIENZA NAVARRO, M.L, *Daños causados por inteligencia artificial y responsabilidad civil*, 2022.
- «¿Una nueva responsabilidad por productos defectuosos? Notas a la Propuesta de Directiva del Parlamento Europeo y del Consejo sobre responsabilidad por daños causados por productos defectuosos de 28 de septiembre de 2022 (COM/2022/495)», *Indret*, 2, 2023, pp. 1-53.
- «Distintos modelos para la responsabilidad civil por daños causados por la Inteligencia Artificial. A modo de comparación entre distintas propuestas de la Unión Europea», en M. Peguera Poch (coord.), *Perspectivas regulatorias de la inteligencia artificial en la Unión Europea*, Madrid, 2023, pp. 295-342.
- «La responsabilidad civil por daños causados por inteligencia artificial. Estado de la cuestión», en N. Álvarez Lata (coord.) /J.M. Busto Lago (pr.), *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, Madrid, 2024, pp. 341-409.
- BARRERO ORTEGA, A., «La tutela antidiscriminatoria en la jurisprudencia del tribunal de justicia de la unión europea», en J.M. Morales Ortega (dir.), *Realidad social y discriminación. Estudios sobre diversidad e inclusión laboral*, Murcia, 2022, pp. 43-57.

- BARRIOS ANDRÉS, M., *Manual de Derecho Digital*, Valencia, 2020.
- BARROS VALE, S./ZANFIR-FORTUNA, G., *Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities*, Future of Privacy Forum, 2022.
- BATTAGLINI MANRIQUE DE LARA, M., «Sentencia histórica del Tribunal de la Haya anulando la elaboración de perfiles para el fraude de la Seguridad Social (SyRI)», accesible en <https://www.linkedin.com/pulse/sentencia-hist%C3%B3rica-del-tribunal-de-la-haya-anulando-manuela/>
- «Cómo el Código Postal sirve para la discriminación de precios», 18 de enero de 2020, accesible en <https://www.linkedin.com/pulse/c%C3%B3mo-el-c%C3%B3digo-postal-sirve-para-la-discriminaci%C3%B3n-de-manuela/>
- BAUZÁ REILLY, M. (coord.)/COTINO HUESO, L. (dir.), *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Thomson Reuters Aranzadi, Navarra, 2022.
- BAZ LOMBA, C., «Los algoritmos y la toma de decisiones administrativas. Especial referencia a la transparencia», *Revista CEFLegal*, 243, 2021, pp. 119-160.
- BELTRÁN DE HEREDIA, I., «Automatización y obsolescencia humana», en *Retos jurídicos de la inteligencia artificial*, coord. por Agustí Cerrillo i Martínez y Miquel Peguera Poch, 2020, pp. 113-124.
- BERTOLINI, A., «Artificial Intelligence and Civil Liability», *Study Requested by the European Parliament's Committee on Legal Affairs*, 2020.
- BERTOLINI, A./EPISCOPO, F., «The Expert Group's Report on Liability for Artificial Intelligence and Other Emerging Digital Technologies: a critical assessment», *European Journal of Risk Regulation*, Vol. 12:3, 2021, pp. 644-659. DOI: <https://doi.org/10.1017/err.2021.30>
- BIBAL A./LOGNOUL, M./DE STREEL, A./FRÉNAY, B., «Legal requirements on explainability in machine learning», *Artificial Intelligence & Law*, 29(2), 2021, pp. 149-169.
- BUOLAMWINI, J./GEBRU, T. «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification», Conference on Fairness, Accountability and Transparency, *Proceedings of Machine Learning Research*, vol. 81, p. 1-15. Disponible en: <https://proceedings.mlr.press/v81/buolamwini18a.html>
- CANTERO ÁLVAREZ, H., *Justicia algorítmica: el sesgo de los algoritmos*, Universidad Politécnica de Madrid, 2021.
- CAPDEFERRO VILLAGRASA, O., «La inteligencia artificial en el sector público: desarrollo y regulación de la actuación administrativa inteligente en la cuarta revolución industrial», *Revista de internet, derecho y política*, n.º 30, Marzo, 2020, pp. 1-14.
- CASADESÚS RIPOLL, P., *La responsabilidad civil por el uso discriminatorio de los datos personales a través de la inteligencia artificial*, Granada, 2022.
- CASTAÑEDA, J. Y OTROS, «Dealing with Gender Bias Issues in Data-Algorithmic Processes: A Social-Statistical Perspective», *Algorithms* 2022, 15, 303. <https://doi.org/10.3390/alg15030303>

- [tps:// doi.org/10.3390/a15090303](https://doi.org/10.3390/a15090303). Recuperado de <https://www.uoc.edu/es/news/2022/297-algoritmos-genero>
- CASTELLANOS CLARAMUNT, J., «Derecho e inteligencia artificial: atención especial a los sesgos, la privacidad y la protección de datos», accesible en <https://idpbarcelona.net/derecho-e-inteligencia-artificial-atencion-especial-a-los-sesgos-la-privacidad-y-la-proteccion-de-datos/>
- CASTILLA, K., *Cuatro ángulos de análisis de la igualdad y la no discriminación en la inteligencia artificial*, Barcelona, 2022.
- CASTILLO OCARANZA, C., «Discriminación algorítmica en el ámbito laboral», en P. Rivas Vallejo (dir.), *Discriminación algorítmica en el ámbito laboral*, Navarra 2022, pp. 71-78.
- CASTILLO PARRILLA, J.A., «Sentencia del Tribunal Ordinario de Bolonia de 31 de diciembre de 2020 (Caso Deliveroo) ¿Discriminación algorítmica o discriminación a través de un algoritmo?», *Derecho Digital e Innovación. Digital Law and Innovation Review*, n.7 (octubre-diciembre), 2020.
- CHEONG, M., Y OTROS, «Ethical Implications of AI Bias as a Result of Workforce Gender Imbalance», CIS & The Policy Lab, The University of Melbourne.
- CHOCRÓN GIRÁLDEZ, A.M., «La acción civil para la defensa de los derechos derivados de la igualdad de trato y la no discriminación», *Indret: Revista para el Análisis del Derecho*, 4, 2023, pp. 83-109.
- COBREROS, L./ GALINDO, J./ RAIGADA, T., *Mujeres en STEM: Desde la educación básica hasta la carrera laboral*, ESADE, Barcelona, 2024.
- COCA PAYERAS, M., «Las iniciativas de la Unión Europea sobre inteligencia artificial: De la persona electrónica al difícil equilibrio entre la necesidad de impulsarla y evitar sus riesgos», *Revista de Derecho Civil*, Vol. 10, n. 2, 2023, pp. 3-40.
- COTINO HUESO, L., «Riesgos e impactos del Big Data, la inteligencia artificial y la robótica. enfoques, modelos y principios de la respuesta del derecho», *Revista General de Derecho Administrativo*, N° 50, 2019.
- «Holanda: “SyRI, ¿a quién sanciono?” Garantías frente al uso de inteligencia artificial y decisiones automatizadas en el sector público y la sentencia holandesa de febrero de 2020”, *La Ley privacidad*, N° 4 (Abril-junio 2020), 2020 (Ejemplar dedicado a: Lecciones aprendidas desde el confinamiento).
 - «Un análisis crítico constructivo de la Propuesta de Reglamento de la Unión Europea por el que se establecen normas armonizadas sobre la Inteligencia Artificial (Artificial Intelligence Act)», *Diario La Ley*, 2 de julio, 2021.
 - «Transparencia de la inteligencia artificial pública: marco legal, desafíos y propuestas», *Actualidad Administrativa*, N° I, *Sección Actualidad*, Diciembre 2023.
 - «Discriminación, sesgos e igualdad de la inteligencia artificial en el sector público», en GAMERO CASADO/PÉREZ GUERRERO (coords.), *Inteligencia artificial y sector público: retos, límites y medios* Valencia, 2023, pp. 260-353.

- «Caso Bosco, a la tercera tampoco va la vencida. Mal camino en el acceso los algoritmos públicos», *Diario LA LEY*, N° 84, *Sección Ciberderecho*, 17 de Mayo de 2024.
 - «La primera sentencia del Tribunal de Justicia de la Unión Europea sobre decisiones automatizadas y sus implicaciones para la protección de datos y el Reglamento de inteligencia artificial» *Diario LA LEY*, N° 80, *Sección Ciberderecho*, 17 de Enero de 2024.
- CRESPO ORTIZ, D., «La indemnización adicional por vulneración de derechos fundamentales del Artículo 183 LRJS. Reflexiones en torno a sus pretensiones y su cálculo», *Labos*, Vol. 4, No. 3, pp. 156-180.
- DE LA CUEVA, J., «Código fuente, algoritmos y fuentes del derecho», *El Notario del Siglo XXI*, núm. 77, 2018. pp. 36-39.
- «El derecho a no ser gobernados mediante algoritmos secretos», *El Notario del Siglo XXI*, núm. 87, 2019.
- DE LA NUEZ SÁNCHEZ CASCADO, E., «Algoritmos y transparencia», disponible en <https://www.hayderecho.com/2020/02/18/algoritmos-y-transparencia-2/>
- DÍAZ ALABART, S., «La responsabilidad civil por discriminación de las personas por causa de su discapacidad (Art. 18 Liondap)», en J.A. Moreno Martínez (coord.), *La responsabilidad civil y su problemática actual*, Madrid, 2007, pp. 181-204.
- *Robots y responsabilidad civil*, Madrid, 2018.
- DIGITAL FUTURE SOCIETY, *La discriminación algorítmica en España: límites y potencial del marco legal*, 2022.
- EBERS, M., «La utilización de agentes electrónicos inteligentes en el tráfico jurídico: ¿Necesitamos reglas especiales en el Derecho de la responsabilidad civil?», *InDret*, n. 3, 2016.
- EBERS, M./ HOCH, V.R.S./ ROSENKRANZ, F./ RUSCHEMEIER, H./ STEINRÖTTER, B., «The European Commission's Proposal for an Artificial Intelligence Act - A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)», *J—Multidisciplinary Scientific Journal*, 2021, 4, pp. 589–603. <https://doi.org/10.3390/j4040043>.
- EGUÍLUZ CASTAÑEIRA, J.A., «Desafíos y retos que plantean las decisiones automatizadas y los perfilados para los derechos fundamentales», *Estudios de Deusto: revista de Derecho Público*, vol. 68, n. 2, 2020, pp. 325-367.
- ERCILLA GARCÍA, J., *Normas de Derecho Civil y Robótica. Robots inteligentes, personalidad jurídica, responsabilidad civil y regulación*, Navarra, 2018.
- ESPÍN ALBA, I., «Sesgos discriminatorios en la toma automatizada de decisiones en la contratación y protección de datos» en N. Álvarez Lata (coord.) /J.M. Busto Lago (pr.), *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, Madrid, 2024, pp. 123-186.

- ESPINOZA VALENCIA, F.F., «Ese algoritmo, el discriminador», recuperado de <https://www.abogacia.es/publicaciones/blogs/blog-de-innovacion-legal/ese-algoritmo-el-discriminador/>
- EUBANKS, V., *Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor*, St Martin's Press, 2018.
- EVANGELIO LLORCA, R., «Causalidad y responsabilidad civil por daños ocasionados por sistemas de inteligencia artificial: las presunciones de causalidad en las propuestas normativas de la UE», en N. Álvarez Lata (coord.) /J.M. Busto Lago (pr.), *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, Madrid, 2024, pp. 549-619.
- FERNÁNDEZ, C., «Primera sentencia europea que declara ilegal un algoritmo de evaluación de características personales de los ciudadanos», Wolters Kluwer a 13 de febrero de 2020. Recuperado de <https://diariolaley.laleynext.es>
- FERNÁNDEZ ARIENZA, A., «Comentario a la sentencia la Corte de apelaciones en el caso "State of Wisconsin v. Eric L. LOOMIS" del Tribunal Supremo del Estado de Wisconsin», *Derecho Digital e Innovación*, n° 5, abril-junio, 2020.
- FERNÁNDEZ DE LA MORENA, B., *Discriminación algorítmica. Estudio del sesgo en arquitecturas de aprendizaje profundo*, Madrid, 2019.
- FERNÁNDEZ HERNÁNDEZ, C., «Modelos para identificar y gestionar los sesgos en la Inteligencia Artificial», *Diario La Ley*, N° 60, Sección Ciberderecho, 21 de Marzo de 2022.
- FERNÁNDEZ ROZAS, J.C., «La Ley de Inteligencia Artificial de la Unión Europea: un modelo para innovaciones radicales, responsables y transparentes basadas en el riesgo», *LA LEY Unión Europea*, N° 124, Sección Estudios, Abril 2024, LA LEY 15196/2024.
- FERNÁNDEZ SÁNCHEZ, S., «Frank, el algoritmo consciente de Deliveroo. Comentario a la Sentencia del Tribunal de Bolonia 2949/2020, de 31 de diciembre», *Revista de Trabajo y Seguridad Social. CEF* n. 457, pp. 179-193.
- GARCÍA GARCÍA, S., «Una aproximación a la futura regulación de la inteligencia artificial en la Unión Europea», *Revista de Estudios Europeos*, n. 79, 2022, pp. 304-323.
- GARCÍA RUBIO, M.P., «La igualdad de trato entre hombres y mujeres y su repercusión en el derecho de contratos. Análisis del Proyecto de Ley Orgánica para la igualdad de mujeres y hombres a la luz de la Directiva 2004/113/CE», *Diario La Ley*, N° 6602, Sección Unión Europea, 30 de Noviembre de 2006, Año XXVII, Ref. D-256 (LA LEY 4083/2006).
- «Discriminación por razón de sexo y Derecho contractual en la Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva de mujeres y hombres», *Derecho Privado y Constitución*, n. 21, 2007, pp. 131 y ss.
- GARCÍA TERUEL, R.M., «El derecho de daños ante la inteligencia artificial y el *machine learning*: una aproximación desde las recomendaciones del Parlamento Europeo y del Grupo de Expertos de la Comisión Europea», en *Cuestiones clásicas*

- cas y actuales del Derecho de daños: estudios en homenaje al profesor Dr. Roca Guillamón*, coord. por Joaquín Ataz López y José Antonio Cobacho Gómez, Vol. 2, 2021, pp. 1009-1055.
- GASCÓN MARCÉN, A., «Derechos humanos e inteligencia artificial», en *Setenta años de Constitución Italiana y cuarenta años de Constitución Española*, Pérez Miras (dir.), Teruel Lozano (dir.), Raffiotta (dir.), Pia Iadicco (dir.), Vol. 5, 2020 (Retos en el siglo XXI / coord. por Silvia Romboli), pp. 335-350.
- GERARDS, J./ XENIDIS, R., *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law* (Special report European network of legal experts in gender equality and non-discrimination), European Commission, 2020.
- GIL GONZÁLEZ, E., *Big Data, Privacidad Y Protección De Datos*, Accedit 2015 al Premio de Protección de Datos Personales de Investigación de la Agencia Española de Protección de Datos, 2016.
- GIL MEMBRADO, C., «En el horizonte: la Directiva de responsabilidad civil extracontractual en materia de inteligencia artificial», en AA. VV., *Derecho y Medicina, desafíos tecnológicos y científicos*, (dir. C. GIL MEMBRADO), Madrid, 2023, pp.101-128.
- «Daños producidos por la IA: la opacidad del algoritmo y el efecto de caja negra», en N. Álvarez Lata (coord.) /J.M. Busto Lago (pr.), *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, Madrid, 2024, pp. 501-547.
- GINÈS i FABRELLAS, A., «Sesgos discriminatorios en la automatización de decisiones en el ámbito laboral: evidencias de la práctica», en RIVAS VALLEJO, P. (dir), *Discriminación algorítmica en el ámbito laboral*, Navarra, 2022, pp. 295-332.
- GÓMEZ LIGÜERRE, C.I., «La Propuesta de Directiva sobre responsabilidad por daños causados por productos defectuosos», *Indret*, n. 4, 2022, pp. 1-7.
- «Concepto de daño moral», en Gómez Pomar, F./Marín García, I. (coords.), *El daño moral y su cuantificación*, Madrid, 2023, pp. 29-72.
- GONZÁLEZ VALVERDE, A., «Responsabilidad por el potencial de sesgo discriminatorio de los algoritmos de los productos de Inteligencia Artificial. A propósito de la Algorithmic Accountability Act of 2019 (s.1108)», en Ataz López y Cobacho Gómez, *Cuestiones clásicas y actuales del Derecho de daños: Estudios en homenaje al profesor Dr. Roca Guillamón*, Vol. 2, 2021, pp. 1225-1273.
- GOODMAN, B/FLAXMAN, S., «European Union regulations on algorithmic decision-making and a “right to explanation”», presented at *ICML Workshop on Human Interpretability in Machine Learning*, New York, 2016, <https://arxiv.org/abs/1606.08813v3>
- GRIMALT SERVERA, P., *La responsabilidad civil en el tratamiento automatizado de datos personales*. Granada, 1999.
- HACKER, P., «Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law», 55 *Common*

- Market Law Review*, 2018, pp. 1143-1186. Disponible en <https://ssrn.com/abstract=3164973>
- HERRERA DE LAS HERAS, R., «Protección de datos e inteligencia artificial», en Cruz Blanca/Lledó Benito (coords.), *La robótica y la inteligencia artificial en la nueva era de la revolución industrial 4.0*, 2021, pp. 644-669.
- «El nuevo marco jurídico de la inteligencia artificial: las novedades del Reglamento Europeo», *Actualidad Civil*, n^o 5, *Sección Derecho digital*, Mayo 2024, LA LEY 18067/2024.
- HUERGO LORA, A., «Una aproximación a los algoritmos desde el derecho administrativo», en A. HUERGO LORA (dir.) y G. DÍAZ GONZÁLEZ (coord.), *La regulación de los algoritmos*, Navarra, 2020, pp. 23-87.
- «El proyecto de Reglamento sobre la Inteligencia Artificial», *Almacén del Derecho*, 17 de abril de 2021.
- «Administraciones Públicas e inteligencia artificial: ¿más o menos discrecionalidad?», *El Cronista del Estado Social y Democrático de Derecho*, n^o 96-97 (Octubre-Noviembre), 2021.
- «Por qué aciertan las sentencias sobre el ‘algoritmo’ del bono social eléctrico», *Almacén del Derecho*, 10 de mayo de 2024.
- HUERGO LORA, A. (dir.) / DÍAZ GONZÁLEZ (coord.), *La regulación de los algoritmos*, Navarra, 2020.
- ITURMENDI MORALES, G., *Responsabilidad Civil por Discriminación*, Madrid, 2023.
- JIMÉNEZ HORWITZ, M., «La protección contra la discriminación en las relaciones entre particulares: la evolución desde la responsabilidad extracontractual hasta los remedios por incumplimiento», *ADC*, tomo LXVII, 2014, fasc. II, pp. 475-515.
- JORQUI AZOFRA, M., *Responsabilidad por los daños causados por productos y sistemas de inteligencia artificial*, Madrid, 2023.
- KARNER, E./KOCH, B. /GEISTFELD, M., *Comparative Law Study on Civil Liability for Artificial Intelligence*, estudio encargado por la Comisión Europea, Dirección General de Justicia y Consumidores, al *European Law Institute (ELI)*, Publications Office, Publications Office, 2021, disponible en: <https://data.europa.eu/doi/10.2838/66412>
- LAZCOZ MORATINOS, G., «Análisis jurídico de la toma de decisiones algorítmica en la asistencia sanitaria», en Díaz González, G. (coord.) / Huergo Lora, A., (dir.), *La regulación de los algoritmos*, 2020, pp. 283-297.
- «Análisis de la propuesta de Reglamento sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas», *IUS ET SCIENTIA: Revista electrónica de Derecho y Ciencia*, Vol. 6, N^o. 2, 2020, pp. 26-41.
- LAZCOZ MORATINOS / CASTILLO PARRILLA, «Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: El caso SyRI», *Revista chilena de derecho y tecnología*, vol. 9, n. 1, 2020, pp. 207-225.

- LLORENS ESPADA, J., «Responsabilidades civiles por discriminación por razón de género cuando medie un sistema de inteligencia artificial», en Rivas Vallejo (dir.), *Discriminación algorítmica en el ámbito laboral*, Navarra 2022, pp. 573-602.
- LOUSADA AROCHENA, J.F., «Ley 15/2022, de 12 julio, integral para la igualdad de trato y la no discriminación: incidencia en el Derecho del Trabajo», accesible en <https://elderecho.com/ley-15-2022-igualdad-trato-no-discriminacion-derecho-del-trabajo>
- LÓPEZ-TARRUELLA MARTÍNEZ, A., *Propiedad intelectual e innovación basada en los datos*, Navarra, 2021.
- MANHEIM, K. M., & KAPLAN, L., «Artificial Intelligence: Risks to Privacy and Democracy», *Yale Journal of Law and Technology* 106, vol. 21, 2019, pp. 106-189.
- MARÍN GARCÍA, I./ GÓMEZ POMAR, F. (coords.) *El daño moral y su cuantificación*, Barcelona, 2023.
- MARÍN LÓPEZ, M.J., «La obligación de evaluar la solvencia del prestatario en la Ley 5/2019, de contratos de crédito inmobiliario», *Revista de Derecho Patrimonial*, n. 50/2019, accesible en https://manueljesusmarin.es/wp-content/uploads/2022/02/doctrin_084.pdf
- MARTÍ GRAU, R., «Reflexiones acerca de la Propuesta de Directiva sobre responsabilidad por daños derivados de la inteligencia artificial y su impacto en el Derecho español de daños», *Revista Aranzadi Doctrinal*, n.4/2023, BIB 2023/705.
- MARTÍNEZ, N/ MATUTE, H., «Discriminación racial en la Inteligencia Artificial», <https://ethic.es/2020/09/discriminacion-racial-en-la-inteligencia-artificial-sesgos/>
- MARTÍN-CASALS, M., «Desarrollo tecnológico y responsabilidad extracontractual. A propósito de los sistemas de inteligencia artificial (IA)», en J.M. Pérez Collados (coord.), *La cultura jurídica en la era digital*, Cuadernos Digitales. Derecho y Nuevas Tecnologías, 2022, pp. 101-138.
- «Las propuestas de la Unión Europea para regular la responsabilidad civil por los daños causados por sistemas de inteligencia artificial», *Indret*, 3, 2023, pp. 55-100.
- MAYER-SCHÖNBERGER, V./ CUKIER, K., *Big data, la revolución de los datos masivos*, 2013.
- MARTÍNEZ GUTIÉRREZ, R., «Responsabilidad administrativa por el uso de la Inteligencia Artificial», en Gamero Casado/Pérez Guerrero (coords.), *Inteligencia artificial y sector público: retos, límites y medios* Valencia, 2023, pp. 227-253.
- MARTÍNEZ, N./MATUTE, H., «Discriminación racial en la Inteligencia Artificial», 2020. Recuperado de <https://ethic.es/2020/09/discriminacion-racial-en-la-inteligencia-artificial-sesgos/>
- MEDINA GUERRERO, M., «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*, n° 49, 2022, pp. 141-171.

- MERCADER UGUINA, J.R. «Discriminación algorítmica y derecho granular: nuevos retos para la igualdad en la era del big data», *Labos: Revista de Derecho del Trabajo y Protección Social*, vol. 2, n. 2, 2021, pp. 4-10.
- MEUWESE, A., «Regulating algorithmic decision-making one case at the time. Case note on: District Court of The Hague, 5/02/20, ECLI:NL:RBDHA:2020:865 (NJCM vs the Netherlands (SyRI))», *European Review of Digital Administration & Law*, 1(1), 2020, pp. 209-211.
- MINISTERIO DE TRABAJO Y ECONOMÍA SOCIAL, *Información algorítmica en el ámbito laboral. Guía Práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral*, mayo 2022.
- MORDUCHOWICZ, R., *La inteligencia artificial ¿Necesitamos una nueva educación?*, UNESCO, 2023.
- MORENO MARTÍNEZ, J.A., «El impacto del Reglamento General de Protección de Datos en el régimen de responsabilidad civil (art. 82 RGPD): Su posible desarrollo por el Derecho interno y problemática de coexistencia con otros mecanismos protectores», en Ataz López, J. y Cobacho Gómez, J.A. (coords.), *Cuestiones clásicas y actuales del Derecho de daños: estudios en homenaje al profesor Dr. Roca Guillamón*, Tomo 3, 2021, pp. 515-565.
- MUÑOZ GARCÍA, C., «Adaptar o reformular la directiva 85/374 sobre responsabilidad por daños causados por productos defectuosos a la inteligencia artificial», *Revista Crítica de Derecho Inmobiliario*, Año 98, n° 793, 2022, pp. 2886-2908.
- MUÑOZ GUTIÉRREZ, C., «La discriminación en una sociedad automatizada: Contribuciones desde América Latina», *Revista chilena de derecho y tecnología*, 10(1), 2021, pp. 271-307.
- MUÑOZ RUIZ, A.B., «¿Se deben regular los algoritmos? Un breve análisis a la propuesta normativa alemana: la pirámide de criticidad basada en el riesgo», recuperado de <https://www.elforodelabos.es/se-deben-regular-los-algoritmos-un-breve-analisis-a-la-propuesta-normativa-alemana-la-piramide-de-criticidad-basada-en-el-riesgo/>
- MUÑOZ VELA, J.M., *Derecho de la Inteligencia Artificial. Un enfoque global de responsabilidad desde la ética, la seguridad y las nuevas propuestas reguladoras europeas*, Tesis Doctoral, Valencia, 2021.
- NAGY, C. I., *Collective Actions in Europe: A Comparative, Economic and Transsystemic Analysis*, Cham, Springer, 2019.
- NAVAS NAVARRO, S., «Negativa a contratar y prohibición de discriminar (Derecho comunitario y Derecho español)», *ADC*, tomo LX, 2007, fasc. IV, pp. 1619 y ss.
- «Sistemas expertos basados en inteligencia artificial y responsabilidad civil. Algunas cuestiones controvertidas», *Diario La Ley, Ciberderecho*, 2019, pp. 1-11 (LA LEY 14980/2019).
- «Responsabilidad civil del fabricante y tecnología inteligente. Una mirada al futuro», *Diario La Ley*, N° 35, Sección Ciberderecho, 27 de Diciembre de 2019, pp. 1-17 (LA LEY 15349/2019).

- «La perspectiva de género en la inteligencia artificial», *Diario La Ley*, n.º 48, *Sección Ciberderecho*, 8 de marzo de 2021.
- *Daños ocasionados por sistemas de inteligencia artificial. Especial atención a su futura regulación*, Granada, 2022.
- «Régimen europeo en ciernes en materia de responsabilidad derivada de los Sistemas de Inteligencia Artificial», *Revista CESCO de Derecho de Consumo*, n. 44, 2022, pp. 27-51.
- «‘Deshumanizando’ al contrato: una mirada desde la inteligencia artificial», en N. Álvarez Lata (coord.) / J.M. Busto Lago (pr.), *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, Madrid, 2024, pp. 31-100.
- NUÑEZ SEOANE, J., «El derecho a la información y acceso al funcionamiento de los algoritmos que tratan datos personales», en *La regulación de los algoritmos*, coord. por Gustavo Manuel Díaz González y Alejandro José Huergo Lora (dir.) 2020, pp. 299-315.
- NUÑEZ ZORRILLA, M.C., «Los nuevos retos de la Unión Europea en la regulación de la responsabilidad civil por los daños causados por la inteligencia artificial», *Revista española de derecho europeo*, núm. 66, 2018, pp. 9-53.
- O’NEIL, C., *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, 2017.
- ORMAZABAL SÁNCHEZ, G., *Discriminación y carga de la prueba en el proceso civil*, Madrid, 2011.
- ORTEGA BALANZA, M./ PÉREZ MUÑOZ, E., «Titulaciones informáticas: equidad e integración de la perspectiva de género en el diseño de algoritmos», en Rivas Vallejo (dir.), *Discriminación algorítmica en el ámbito laboral*, 2022, pp. 669-692.
- ORTIZ FERNÁNDEZ, M., *La responsabilidad civil derivada de los daños causados por sistemas inteligentes y su aseguramiento: análisis del tratamiento ofrecido por la Unión Europea*, Madrid 2021.
- ORWAT, C., *Risks of Discrimination through the Use of Algorithms. A study compiled with a grant from the Federal Anti-Discrimination Agency*, Berlín, 2020, recuperado https://www.antidiskriminierungsstelle.de/EN/homepage/_documents/download_diskr_risiken_verwendung_von_algorithmen.pdf?__blob=publicationFile&v=1
- OTTOLIA, A., *Derecho, Biga Bata e Inteligencia Artificial*, Valencia, 2018.
- PALMA ORTIGOSA, A., «Decisiones automatizadas en el RGPD. El uso de algoritmos en el contexto de la protección de datos», *Revista General de Derecho Administrativo*, n. 50, 2019.
- «El ciclo de vida de los sistemas de inteligencia artificial. Aproximación técnica de las fases presentes durante el diseño y despliegue de los sistemas algorítmicos», en *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, M. Bauzá Reilly (coord.); L. Cotino Hueso (dir.), Navarra, 2022, pp. 29 y ss.
- PARLAMENTO EUROPEO, *The ethics of artificial intelligence: Issues and initiatives* (Study - Panel for the Future of Science and Technology), marzo 2020.

Recuperado de [https://www.europarl.europa.eu/thinktank/es/document/EPRS_STU\(2020\)634452](https://www.europarl.europa.eu/thinktank/es/document/EPRS_STU(2020)634452)

- PEGUERA POCH, M., «La propuesta de Reglamento de la IA: una intervención legislativa insoslayable en contexto de incertidumbre», en M. Peguera Poch (coord.) *Perspectivas regulatorias de la inteligencia artificial en la Unión Europea*, Madrid, 2023, pp. 125-182.
- «Perspectiva y categorización del riesgo en el Reglamento de Inteligencia Artificial», en J.A. Moreno Martínez/P.J. Femenía López (coords.), *Inteligencia artificial y derecho de daños: Cuestiones actuales: Acorde al Reglamento (UE) 2024/1689*, Madrid, 2024, pp. 923-948.
- PELÁEZ VARGAS, G., «Indicios y presunciones», *Revista Facultad de Derecho y Ciencias Políticas*, n. 48, 1974, pp. 49-72.
- PEÑA LÓPEZ, F., *La culpabilidad en la responsabilidad civil extracontractual*, Granada, 2002.
- «Responsabilidad objetiva y subjetiva en las propuestas legislativas europeas sobre responsabilidad civil aplicables a la inteligencia artificial», en N. Álvarez Lata (coord.) /J.M. Busto Lago (pr.), *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, Madrid, 2024, pp. 411-499.
- «¿Qué daño moral se presume que se deriva de los actos de discriminación y qué tipo de presunción establece la Ley 15/2022?», *CESCO*, septiembre 2022, pp. 1-8.
- PLAZA PENADES, J., «Los contratos informáticos y electrónicos», en J.R. De Verda y Beamonte (coord.) *Derecho civil II. Obligaciones y contratos*, Valencia, 2021, pp. 567-581.
- RAMÍREZ-BUSTAMANTE, N./ PÁEZ, A., «Análisis jurídico de la discriminación algorítmica en los procesos de selección laboral», 2021. Disponible en <https://ssrn.com/abstract=3765741> o <http://dx.doi.org/10.2139/ssrn.3765741>
- REBOLLO DELGADO, L. *Inteligencia artificial y derechos fundamentales*, Madrid, 2023.
- REY MARTÍNEZ, F., *Derecho Antidiscriminatorio*, Navarra, 2019.
- RIBES RIBES, A., «La posición del contribuyente ante los sistemas de inteligencia artificial utilizados por la Administración», *Revista Aranzadi*, 2021.
- RIVAS VALLEJO, P., «Sesgos de automatización y discriminación algorítmica», en *Discriminación algorítmica en el ámbito laboral*, Pilar Rivas Vallejo (dir.), 2022, pp. 31-68.
- «Análisis desde el derecho antidiscriminatorio», en P. Rivas Vallejo (dir.) *Discriminación algorítmica en el ámbito laboral*, Navarra 2022, pp.415-476.
- «Herramientas desde el derecho antidiscriminatorio y la protección frente a la igualdad y a la no discriminación», en P. Rivas Vallejo (dir.) *Discriminación algorítmica en el ámbito laboral*, Navarra 2022, pp. 539-572.
- RODRÍGUEZ ÁLVAREZ, A., «La carga de la prueba en la Ley integral para la igualdad de trato y la no discriminación: algunas consideraciones críticas desde la óptica

- del proceso civil», *Diario La Ley*, N° 10131, *Sección Tribuna*, 15 de septiembre de 2022 (LA LEY 7764/2022).
- RODRÍGUEZ CARDO, I.A., «Decisiones automatizadas y discriminación algorítmica en la relación laboral: ¿hacia un Derecho del Trabajo de dos velocidades?», *Revista Española de Derecho del Trabajo* num.253/2022, BIB 2022\1779.
- ROIG, A., *Las garantías frente a las decisiones automatizadas: del Reglamento General de Protección de Datos a la gobernanza algorítmica*, Barcelona, 2020.
- RUBÍ PUIG, A., «Problemas de coordinación y compatibilidad entre la acción indemnizatoria del artículo 82 del Reglamento General de Protección de datos y otras acciones en derecho español», *Derecho Privado y Constitución*, n. 34, 2019, pp. 197-232.
- «Retos de la inteligencia artificial y adaptabilidad del derecho de daños», *Retos jurídicos de la inteligencia artificial*, coord. por Agustí Cerrillo i Martínez y Miquel Peguera Poch, 2020, pp. 57-71.
- «Elaboración de perfiles y personalización de ofertas y precios en la contratación con consumidores», *Revista De Educación Y Derecho*, 24, 2021.
- «Responsabilidad civil e Inteligencia artificial. Un examen crítico de la propuesta de Directiva de 28 de septiembre de 2022», en M. Peguera Poch (coord.) *Perspectivas regulatorias de la inteligencia artificial en la Unión Europea*, Madrid, 2023, pp. 245-294.
- «Inteligencia artificial y daños indemnizables», en N. Álvarez Lata (coord.) /J.M. Busto Lago (pr.), *Derecho de contratos, responsabilidad extracontractual e inteligencia artificial*, Madrid, 2024, pp. 621-688.
- SÁEZ LARA, C., «El algoritmo como protagonista de la relación laboral. Un análisis desde la perspectiva de la prohibición de discriminación», *Temas laborales: Revista andaluza de trabajo y bienestar social*, N° 155, 2020, pp. 41-60.
- SALAZAR GARCÍA, I., «Retos actuales de la ética en la inteligencia artificial», en *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, coord. por Marcelo Bauzá Reilly; Lorenzo Cotino Hueso (dir.), 2022, pp. 53-66.
- SÁNCHEZ CAPARRÓS, M. «Prevenir y controlar la discriminación algorítmica», *Rubinzal Culzoni*, n° 427, 2021. Recuperado de https://www.researchgate.net/publication/358207305_Prevenir_y_controlar_la_discriminacion_algoritmica
- SCHEUERMAN/PAU/BRUBAKER, «How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services», *Proc. ACM Hum. Comput. Interact.* 3, CSCW, Article 144, 2019, <https://doi.org/10.1145/3359246>
- SELBST, A. D./POWLES, J., «Meaningful information and the right to explanation», *International Data Privacy Law*, 2017, Vol. 7, n.º 4, pp. 233 y ss.
- SOLAR CAYÓN, J.I., «Inteligencia artificial en la justicia penal: los sistemas algorítmicos de evaluación de riesgos», en *Dimensiones éticas y jurídicas de la inteligencia artificial en el marco del Estado de Derecho*, J.I. Solar Cayón (ed.), Alcalá de Henares, 2020, pp. 125-172.

SORIANO ARNANZ, A. «Decisiones automatizadas y discriminación: aproximación y propuestas generales», *Revista General de Derecho Administrativo*, n. 56, 2021.

— «Decisiones automatizadas: problemas y soluciones jurídicas. Más allá de la protección de datos», *Revista de Derecho Público: teoría y método*, n. 13, 2021, pp. 85-127.

— «Discriminación algorítmica: garantías y protección jurídica», en *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, coord. por M. Bauzá Reilly; dir. por L. Cotino Hueso, 2022, pp. 139-169.

— «La aplicación del marco jurídico europeo en materia de igualdad y no discriminación al uso de aplicaciones de Inteligencia Artificial», en *Nuevas normativas: inteligencia artificial, derecho y género* (coord. por P.R. Bonorino Ramírez / R. Fernández Acevedo / P. Valcárcel Fernández), Navarra, 2022, pp. 63-88.

— «La Propuesta de Reglamento de Inteligencia Artificial de la UE y los Sistemas de Alto Riesgo», *Revista General de Derecho de los Sectores Regulados*, n.8, noviembre 2021, pp.1-24.

SORIANO ARNANZ, A./SIMÓ SOLER, E., «Machine learning y Derecho: aprendiendo la (des)igualdad», en S. Barona Vilar (ed.) *Justicia algorítmica y neuroderecho: una mirada multidisciplinar*, Valencia, 2021, pp. 183-207.

SUBERBIOLA GARBIZU, I., «Inteligencia artificial, opacidad y motivación de las decisiones administrativas automatizadas», en *La protección de los derechos fundamentales en el ámbito tributario*, coord. por Antonio Vázquez del Rey Villanueva e Irune Suberbiola Garbizu y dir., por Isaac Merino Jara, 2021, pp. 291-328.

SUNÉ LLINÁS, E., *Derecho e Inteligencia Artificial. De la robótica a lo posthumano*, Valencia, 2020.

SWEENEY, L., «Discrimination in Online Ad Delivery», 2013. Disponible en <http://dx.doi.org/10.2139/ssrn.2208240>

TAPIA HERMIDA, A., «Propuesta de Directiva sobre responsabilidad en Inteligencia Artificial Responsable (IAR)», en <https://sanchezcalero.com/propuesta-de-directiva-sobre-responsabilidad-en-inteligencia-artificial-responsable-iar/>

— «La responsabilidad civil derivada del uso de la inteligencia artificial y su aseguramiento», *Revista de la Asociación Española de Abogados Especializados en Responsabilidad Civil y Seguro*, n. 76, 2020, pp. 79-104.

TIMÓN HERRERO, M., «La transparencia de la actuación de las Administraciones Públicas. Cinco desafíos y algunas propuestas», recuperado de <https://elderecho.com/transparencia-actuacion-de-administraciones-publicas>

TOLOSA, P./DIBO, C., «Inteligencia artificial, discriminación por género y Derecho: viejos problemas, nuevos desafíos», en C. Danesi (dir.) *Inteligencia Artificial, tecnologías emergentes y Derecho*, 2021, pp. 155-190.

TORRALBA ROSELLÓ, F. Y OTROS, «Ètica algorítmica i perspectiva de gènere: de l'opacitat a la transparència», *Ars Brevis: anuario de la Càtedra Ramon Llull Blanquerna*, n. 27, 2021, pp. 211-282

- VANTIN, S. «Inteligencia artificial y derecho antidiscriminatorio», en *Inteligencia Artificial y Derecho. El jurista ante los retos de la era digital*, coord. por Fernando Higinio Llano Alonso, Joaquín Garrido Martín, 2021, pp. 367-384.
- VÁZQUEZ DE CASTRO, E., «Aproximación a la responsabilidad derivada de los riesgos de la inteligencia artificial en Europa», en *Dimensiones éticas y jurídicas de la inteligencia artificial en el marco del Estado de Derecho* J.I., Solar Cayón (ed. lit.), 2020, pp. 229-273.
- VELASCO RICO, C., «Marco regulatorio de los sistemas algorítmicos y de IA. El papel de la Administración», ponencia en el XVIII Congreso de la Asociación Española de Profesores de Derecho Administrativo: El derecho administrativo en la era de la inteligencia artificial, Vigo 25 a 27 enero de 2024.
- VESTRI, G., «La inteligencia artificial ante el desafío de la transparencia algorítmica. Una aproximación desde la perspectiva jurídico-administrativa», *Revista Aragonesa de Administración Pública*, n° 56, 2021, pp. 368-398.
- VICENTE PALACIO, M.A., «La inteligencia artificial en la ley integral para la igualdad y no discriminación», *Revista General de Derecho del Trabajo y de la Seguridad Social*, N° 64, 2023, pp.125-150.
- WACHTER, S./MITTELSTADT, B./FLORIDI, L. «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation», *International Data Privacy Law*, Vol. 7, n. 2, 2017, pp. 1-47.
- WASHINGTON, A.L. «How to argue with an algorithm: Lessons from the COMPAS ProPublica Debate», *The Colorado Technology Law Journal*, n° 17, 2018, pp. 131-160.
- WILSON, B./ HOFFMAN, J./MORGENSTERN, J., «Predictive Inequity in Object Detection», 2019 <https://arxiv.org/pdf/1902.11097.pdf>
- XENIDIS, R., «Tuning EU equality law to algorithmic discrimination: Three pathways to resilience», *Maastricht Journal of European and Comparative Law*, n. 27 (6), 2020, pp. 736-758. <https://doi.org/10.1177/1023263X20982173>
- XENIDIS, R./SENDEN, L., «EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination», en Ulf Bernitz et al (eds.), *General Principles of EU law and the EU Digital Order*, Kluwer Law International, 2020.
- YZQUIERDO TOLSADA, M., *Responsabilidad civil extracontractual. Parte general*, Madrid, 2020

GRACIAS POR CONFIAR EN NUESTRAS PUBLICACIONES

Al comprar este libro le damos la posibilidad de consultar gratuitamente la versión ebook.

Cómo acceder al ebook:

- ☞ **Entre en nuestra página web**, sección Acceso ebook
(www.dykinson.com/acceso_ebook)
- ☞ **Rellene el formulario** que encontrará insertando el código de acceso que le facilitamos a continuación así como los datos con los que quiere consultar el libro en el futuro (correo electrónico y contraseña de acceso).
- ☞ Si ya es **cliente registrado**, deberá introducir su **correo electrónico y contraseña habitual**.
- ☞ Una vez registrado, **acceda a la sección Mis e-books de su cuenta de cliente**, donde encontrará la versión electrónica de esta obra ya desbloqueada para su uso.
- ☞ Para consultar el libro en el futuro, ya sólo es necesario que se identifique en nuestra web con su correo electrónico y su contraseña, y que se dirija a la sección Mis ebooks de su cuenta de cliente.



CÓDIGO DE ACCESO

Rasque para ver el código

Nota importante: Sólo está permitido el uso individual y privado de este código de acceso. Está prohibida la puesta a disposición de esta obra a una comunidad de usuarios.



**MANTÉNGASE INFORMADO
DE LAS NUEVAS PUBLICACIONES**

**Suscríbase gratis
al boletín informativo
www.dykinson.com**

Y benefíciase de nuestras ofertas semanales

Son innumerables los sectores y actividades donde se están empleando sistemas de inteligencia artificial que, gracias al procesamiento y combinación de ingentes cantidades de datos personales y no personales, pueden decidir de forma automatizada y con cierto grado de autonomía, por ejemplo, quién merece ser destinatario de un crédito o de una ayuda pública, quién es apto para un determinado puesto de trabajo, quién puede acceder a un centro educativo, o qué probabilidades de reincidencia tiene un preso.

Aunque los beneficios que conlleva esta tecnología son indiscutibles, en ocasiones se ven ensombrecidos por los riesgos que implica y que pasan desapercibidos para los ciudadanos. Además de amenazas a la intimidad o a la protección de datos personales, uno de los mayores peligros del uso de estos sistemas es la adopción de decisiones sesgadas y discriminatorias que lleven a reproducir, amplificar y consolidar estereotipos y desigualdades contra las que se lleva siglos luchando. De hecho, la perpetuación de patrones históricos de discriminación ha sido tenida en cuenta en el reciente *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial*, para clasificar los sistemas inteligentes dentro de uno u otro nivel de riesgo.

La presente monografía aborda los potenciales riesgos y daños al derecho a la igualdad de trato y no discriminación derivados del uso de sistemas de inteligencia artificial en la toma de decisiones. Se estudia, entre otras cuestiones, el concepto y las características de la discriminación algorítmica, las causas que pueden provocarla, los desafíos que plantea y las medidas que pueden adoptarse para hacerles frente. Todo ello mediante el análisis de la legislación sobre igualdad y no discriminación, sobre protección de datos y sobre inteligencia artificial, así como de otras soluciones propuestas desde el ámbito ético y tecnológico para prevenir los riesgos y reparar los daños que este fenómeno puede ocasionar.

