Universidad del País Vasco / Euskal Herriko Unibertsitatea

Facultad de Derecho - Campus de Bizkaia

Departamento de Derecho Público

GOBERNANZA Y SUPERVISIÓN HUMANA DE LA TOMA DE DECISIONES AUTOMATIZADA BASADA EN LA ELABORACIÓN DE PERFILES

-

GOVERNANCE AND HUMAN OVERSIGHT OF AUTOMATED DECISION-MAKING BASED ON PROFILING

Guillermo Lazcoz Moratinos

Doctorando

Prof. Dr. iur. Dr. med. Carlos María Romeo Casabona

Director

2023

Programa de Doctorado en Derechos Humanos, Poderes Públicos, Unión Europea: Derecho Público y Privado

Hemos organizado una civilización global en la que la mayoría de los elementos cruciales (...) dependen profundamente de la ciencia y la tecnología.

También hemos organizado las cosas de manera que casi nadie entiende la ciencia y la tecnología. Esto es una receta para el desastre.

Puede que nos salgamos con la nuestra durante un tiempo, pero tarde o temprano esta mezcla inflamable de ignorancia y poder nos estallará en la cara

Carl Sagan: The Demon-Haunted World: Science as a Candle in the Dark (1997)

Todo se puede hacer en la vida, con método

Lola Flores: El perro verde con Jesús Quintero (1988)

RESUMEN

La toma de decisiones automatizada basada en la elaboración de perfiles que realiza predicciones o clasificaciones sobre nosotros forma ya parte de lo cotidiano. Con la evolución de las tecnologías que permiten este fenómeno aumenta nuestra dependencia de la delegación de tareas en los sistemas algorítmicos y, al mismo tiempo, se genera una expectativa de confianza sobre dicha delegación. Esta delegación se traduce también en una reducción gradual de la participación humana de muchos de estos procesos de toma de decisiones. No obstante, el Derecho no parece querer renunciar a la supervisión humana de estos procesos, especialmente cuando esta toma de decisiones produce efectos significativos en las personas sobre las que se realizan predicciones o clasificaciones.

Siguiendo esta misma lógica, el RGPD prohíbe en su artículo 22 la toma de decisiones, incluida la elaboración de perfiles, basada únicamente en el tratamiento automatizado, imponiendo la inclusión de intervención humana en el proceso decisorio para esquivar dicha prohibición. Incluso cuando excepcionalmente habilita la toma de decisiones basada únicamente en el tratamiento automatizado impone una intervención humana posterior a la adopción de la decisión como medida de salvaguarda para los derechos y libertades de las personas interesadas.

Sin embargo, estos mecanismos de gobernanza basados en la intervención humana han recibido escaso interés por la doctrina y la jurisprudencia. Ello genera una notoria falta de seguridad jurídica a la hora de interpretar qué clase de intervención humana es exigida por el RGPD y cómo puede implementarse en un proceso de toma de decisiones automatizada. Esta investigación trata de ser una aportación preliminar en esta línea a partir, primero, del análisis del vilipendiado artículo 22 RGPD; y posteriormente, partiendo del principio nuclear de la responsabilidad en el ecosistema normativo del RGPD, destacando las obligaciones que introduce la evaluación de impacto de protección de datos en relación con la intervención humana para la toma de decisiones automatizada en el RGPD.

A la luz de este análisis jurídico, puede concluirse que el RGPD exige una intervención humana significativa y demostrable para la toma de decisiones automatizada basada en la elaboración de perfiles. Ahora bien, este optimismo debe contrarrestarse con una necesaria dosis de realismo: resulta indispensable -y urgente- introducir reformas para hacer efectiva dicha intervención humana significativa y demostrable. De otro modo, aunque bienintencionados, los mandatos del Derecho para la supervisión humana de estos procesos resultarán inútiles.

ABSTRACT

The use of automated decision-making based on profiling to make predictions or categorize us is becoming commonplace. Our reliance on task delegation to algorithmic systems grows as technology advances, and at the same time, expectations for such delegation are raised. As a result of this delegation, these decision-making processes gradually involve less humans. However, the law does not appear to wish to abandon human control over these procedures, particularly where this decision-making has a major impact on the people that are profiled.

Following the same logic, Article 22 GDPR prohibits decision-making, including profiling, based solely on automated processing, and requires the integration of human intervention in the decision-making process in order to circumvent this prohibition. Even when it exceptionally enables decision-making based solely on automated processing, it imposes human intervention after the decision has been taken as a safeguard for the rights and freedoms of the individuals affected.

However, these governance mechanisms based on human intervention have received little interest among legal scholars and courts. Due to this, there is a well-known lack of legal certainty around the type of human interaction that the GDPR requires and how it can be implemented in an automated decision-making process. This study aims to make a preliminary contribution in this area. First by analysing the Kafkian Article 22 GDPR. Then, focusing on the requirements placed by the data protection impact assessment in respect to human intervention in automated decision-making under the GDPR, we turn to the core principle of accountability in this legal ecosystem.

This legal analysis leads to the conclusion that the GDPR mandates significant and demonstrable human intervention in automated decision-making based on profiling. This optimism, though, needs to be coupled with a robust dose of realism: it is indispensable - and urgent - to introduce amendments to make such meaningful and demonstrable human intervention effective. The law's requirements for human supervision of these procedures would be pointless otherwise, no matter how well-intentioned they may be.

AGRADECIMIENTOS

Empiezo la casa por el tejado.

No puedo dejar de dar las gracias a las personas que forman la Cátedra, en su sentido más amplio, ahora Red Cátedra de Derecho y Genoma Humano. A quienes formáis parte de este grupo tan humano. A mi director y maestro, Romeo Casabona, por ser tan arquitecto como orfebre de este proyecto que nació en 1993. También por desbrozar hábilmente el camino de esta investigación.

A Pilar, por enseñarme a enseñar, porque siempre estás sin necesidad de pedirlo y por tu rigor en todo desempeño. A Iñigo, por tu capacidad para despertar las mentes más letárgicas. Espero que escribas ese esperpéntico guion que comprará la HBO.

A Ángela, por ser una inspiración sin pretenderlo ni lo más mínimo, no sé a quién le toca pagar las cañas. A Mikel, por el día en que tiñamos este mundo de colores más justos. A Daniel y Jose, porque os acogería en mi casa una y mil veces, sois un encanto. A Ekain, Iker y Carlos, porque, acompañada, la penitencia doctoral ha resultado más liviana.

Al alumnado de Sarriko y Leioa por todo el cariño que me brindasteis. Quien critica a quienes venís abriéndoos camino detrás, no tiene ni la más remota idea del tesoro que guardáis.

De mi paso por Bruselas. To Paul, for showing me how to cook good work while having fun. The right place at the right time always comes. A Andrés, por abrirme las puertas de tu hogar siempre que hizo falta. To Anastasiya, for taking me all the way to Siberia with a conversation and a glass of wine. Y a Couronne, por descubrirme una Bruselas tan especial que solo existió en pandemia.

A Carmen, por transmitirme tus incandescentes ganas de aprender de todos y cada una de quienes te rodean. Y aquí de nuevo a Pilar, por querer soportarme un rato más. A ambas, por la forma en la que entendéis el liderazgo.

Y ahora los cimientos.

Quien bien me conoce, bien sabe que para mí este trabajo no significa nada sin estar rodeado de vida. Por eso, a quienes habéis sido, sois y seréis mi vida.

A los Soviet, por haber sido mi casa en Madrid y robarme con cuentos la sangre y la vida de mi corazón. La casa más diversa en la que he habitado y en la que tanto he crecido con las diferencias que nos unen. A Ihani, por enseñarme todo lo que sé, por anclarme mis pecados capitales al suelo para despegar siempre conmigo en esa bohemia y nihilista Malasaña que nos atrapa. A Laura y María, por ser cobijo cuando más lo necesitaba. A Vir, porque todo comenzó en Fuencarral ciento veintisiete.

A Baloncesto Atocha y Ostiko KE por mantener vivo mi amor por el baloncesto durante estos años.

A Aritz, por un corazón que no te cabe en el pecho. A Iván, por darnos las lecciones más valiosas siempre dispuesto a montar jarana. A Mai, por amatxo y liada, sobre todo por liarla juntos. A Marta, por todo lo que nos une. A Julen, porque eres y dejas ser de la forma más bonita que conozco. A David, por no dejarme tener la razón cuando la tengo, y saber dármela cuando no la llevo -pero la necesito-, porque por qué no. A Idoia e Itzi, por poner todo patas arriba perreando hasta el suelo. A Jon y Jorge, porque somos amores distintos y a la vez imposibles de comprender el uno sin el otro, el otro sin aquél y aquél sin el uno. A Josune, por todo lo que ya sabes cuando nos ponemos intensos. A Pati, por reputa y por bonica.

A mi familia por cuidarme y quererme bien.

A mis abuelos Loli y Antonio, por desentrañarnos todos los secretos del amor fraterno a los venideros. A mi tía Loli, por deslumbrarnos y sorprendernos siempre, después de la siesta al bar, al centro. Y mañana a comer, a tu casa.

A los amores de mi vida, Iratxe y Telma. Sois el orgullo que arde y la fuerza mueve el mundo. A Luis, aita, por enseñarme a no agachar la cabeza, a ser ingobernable. A Keles, ama, por enseñarme a ser creativo -aunque para su desdicha nunca desarrollase talento alguno para las bellas artes- y responsable, a partes iguales.

A Adeli, por haber sido mi compañera desde la ternura más alegre, la serenidad y los cuidados. Por bordarme una falda manchega y llana, con tus manos y desde tus raíces.

A las casualidades que no dejan de abrirse paso.

(…)

FINANCIACIÓN

Entre el 30 de octubre de 2017 y el 14 de noviembre de 2021 la presente investigación se realizó gracias a la contratación por la Universidad del País Vasco (UPV/EHU) como Investigador Predoctoral en formación financiada por el Ministerio de Educación, Cultura y Deporte. Dicho contrato fue suscrito como consecuencia de la Convocatoria de becas y ayudas para la formación de doctores del programa nacional de formación de profesorado universitario (FPU) 2016, de acuerdo con las condiciones establecidas en Resolución de 22 de diciembre de 2016 (BOE de 17 de enero de 2017) -FPU16/06314-.

De igual modo, la estancia internacional realizada entre el 4 de enero de 2021 y el 3 de abril de 2021 bajo la supervisión de Paul de Hert en el grupo de investigación Law, Science, Technology & Society (LSTS) de la Vrije Universiteit Brussel (VUB), fue posible gracias a las Ayudas complementarias de movilidad destinadas a beneficiarios del programa de Formación del Profesorado Universitario (FPU) del Ministerio de Universidades (BOE de 15 de junio de 2020) -EST19/00674-.

Como miembro del G.I. Cátedra de Derecho y Genoma Humano de la Universidad del País Vasco (UPV/EHU), entre el 1 de enero de 2018 y el 14 de noviembre de 2021, algunas de las publicaciones y actividades académicas que he realizado en el seno de la presente investigación han contado también con la financiación del Departamento de Educación del Gobierno Vasco para apoyar las actividades de Grupos de Investigación del Sistema Universitario Vasco -IT 1066-16-.

Camarera y camarero en hostelería, ayudante de pastelero, azafata en congresos, peón en cadena industrial, comercial y repartidor en empresas de disolventes, restauradora de obras de arte, autónomo en trabajos verticales, profesora de dibujo y pintura en centros penitenciarios, encargado en empresa de pintura, guarda nocturno, conserjes, limpiadora, perrero municipal, auxiliar de clínica, chófer, jardinería municipal, agente de información y control. Y todo el trabajo no remunerado. Son algunas de las labores desempeñadas por mi madre y mi padre que me han sostenido y permitido disfrutar de una infancia, adolescencia y juventud felices para poder realizar, entre otros, este trabajo de investigación.

TABLA DE CONTENIDOS

PRESENTACIÓN: OBJETO DE ESTUDIO, JUSTIFICACIÓN Y MÉTODO

PRESENTACIÓN: OBJETO DE ESTUDIO, JUSTIFICACIÓN Y MÉTODO

Nuestra sociedad es testigo y parte de un uso extensivo y ubicuo de cada vez más datos -a menudo personales, cuando no sensibles-; junto con una creciente dependencia de algoritmos para analizarlos con el fin de obtener predicciones o clasificaciones -también de carácter personal e incluso sensibles-; para tomar decisiones que afectan a personas físicas. A dicho fenómeno le acompaña una reducción gradual de la participación humana e incluso de la supervisión, también humana, de muchos de estos procesos de toma de decisiones, que plantean cuestiones apremiantes de equidad, responsabilidad y respeto de los derechos humanos, entre otras[1]. Con la evolución y el aumento de la eficacia de las tecnologías de inteligencia artificial (IA en adelante), aumenta la dependencia de la delegación de tareas y, al mismo tiempo, se genera una expectativa de confianza sobre dicha delegación[2].

La irrupción del conjunto de tecnologías involucradas en estas transformaciones, que en los últimos años se han venido a conceptualizar como sistemas algorítmicos o de IA -y a regular como tales-, ha despertado un interés jurídico sin precedentes[3]. Muchas decisiones con importantes implicaciones individuales, colectivas y sociales que antes eran tomadas sólo por humanos -a menudo por expertos y expertas- ahora son tomadas o asistidas por algoritmos o sistemas de IA[4]. Esta investigación, desde una perspectiva jurídica, trata sobre el rol humano en esta clase de decisiones y en las inferencias algorítmicas que sirven de sustento a las mismas.

Ahora bien, la investigación no se centra exclusivamente en las decisiones clave que se delegan a máquinas sin participación humana alguna y, en su caso, en cómo el ordenamiento jurídico responde a esta clase de decisiones y el rol que el ordenamiento otorga al ser humano en estas respuestas. Aunque presenciamos esta clase de automatización "completa" de la toma de decisiones, en la actualidad es aún más habitual

[1] Floridi y Taddeo, «What is data ethics?», 2.

[2] Taddeo, «Trusting Digital Technologies Correctly», 566.

[3] Basta consultar el término "inteligencia artificial" para ciencias jurídicas en el portal bibliográfico Dialnet. Mientras que en la década 2000-2009 arroja un total de 12 resultados, la década 2010-2019 arroja 410. Y en la siguiente década 2020-2029, a fecha 13 de julio de 2022, contiene ya 1.165 resultados. Este crecimiento exponencial pone de manifiesto ese interés jurídico sin precedentes. Sobre cómo se entienden y utilizan términos como algoritmos o sistemas de IA en esta investigación, vid. contextualización técnica en la introducción.

[4] Lepri et al., «Fair, Transparent, and Accountable Algorithmic Decision-making Processes», 612.

que sean personas las que tengan el control "final" sobre la toma de decisiones, si bien, con el apoyo o sustento de clasificaciones -aparentemente- empíricas de riesgo, cuya racionalidad no tienen forma de cuestionar[5]. Ello obliga no solo a prestar atención sobre la delegación de las decisiones en sí misma, sino a cómo se delegan y a quién controlan y cómo influyen los sistemas automatizados en esa delegación[6].

La participación humana y su relevancia jurídica en estos procesos no se limita a la toma de decisiones final. La toma de decisiones automatizada sobre la base de inferencias algorítmicas se comprende de varias fases, fundamentalmente, la fase de diseño y desarrollo del modelo algorítmico y la fase de implementación o despliegue del mismo[7].

La participación humana tiene una relevancia jurídica indiscutible en todas ellas, y no exclusivamente como agente último en la toma de decisiones. Se dice, por ejemplo, que los datos son el combustible de estos algoritmos, no obstante, los datos somos nosotros mismos, las personas y nuestra actividad son en realidad el combustible y materia prima de dichos algoritmos. Los seres humanos se han vuelto detectables, (re)rastreables y correlacionables mucho más allá de su control y los rastros que producen empiezan a vivir su propia vida, convirtiéndose en los recursos de una red muy extensa, si no ilimitada, de posibles dispositivos de elaboración de perfiles que generan conocimientos que les conciernen y/o les afectan directa o indirectamente[8]. Es decir, desde el momento en que estos sistemas perciben e interpretan el entorno, se produce una reordenación de la actividad humana sobre los fundamentos algorítmicos que rigen estas herramientas, y dicha reordenación de la actividad humana tiene su máxima expresión cuando la percepción e interpretación algorítmicas se utilizan como fundamento para la toma de decisiones.

[5] Para McQuillan, la IA fomenta la irreflexibidad en el sentido definido por Arendt, es decir, la incapacidad de criticar instrucciones, la falta de reflexión sobre sus consecuencias y el afán de creer que se está llevando a cabo una ordenación correcta. En este sentido, McQuillan considera que la IA es la burocracia del Siglo XXI. McQuillan, «The Political Affinities of AI», 165.

[6] Djeffal, «AI, Democracy and the Law», 277.

[7] Más adelante se abordarán las sub-fases que comprenden estas dos fases principales como, por ejemplo, la recolección de datos o el entrenamiento dentro de la fase de diseño y desarrollo.

[8] Gutwirth y De Hert, «Regulating Profiling in a Democratic Constitutional State», 291.

«*Las personas están por encima de la tecnología y las máquinas*»[9] dice el artículo 12(1) de la Constitución del *Länder* de Bremen[10], desde su redacción original de 1947. Situar a las personas por encima de la máquina no sólo significa que aquéllas no deben verse perjudicadas por las nuevas posibilidades tecnológicas, sino que deben situarse *en el asiento del conductor*, lo cual puede entenderse como una autodeterminación efectiva de las personas en diferentes niveles[11].

Este mismo principio lo hemos podido ver recogido en posteriores iniciativas para la regulación de la IA en el ámbito europeo. Para el Grupo de expertos de alto nivel sobre inteligencia artificial de la Comisión Europea (HLEG-AI en adelante), la autodeterminación humana respecto del desarrollo tecnológico es el fundamento clave para la reivindicación de una IA centrada en el ser humano y, en particular, para la inclusión de la acción y supervisión humanas como uno de los siete requerimientos para el desarrollo de una IA fiable[12]. Asimismo, en el año 2020 la supervisión humana fue recogida también como requerimiento de obligado cumplimiento para las aplicaciones de IA de alto riesgo en el Libro Blanco de la Comisión Europea[13] e igualmente, para cumplir con la exigencia del desarrollo de una IA antropocéntrica y antropogénica, en la

[9] Del original en alemán: "Der Mensch steht höher als Technik und Maschine".

[10] No está de más puntualizar que los *Länder* alemanes disponen de autonomía constitucional pudiendo, por ende, darse una Constitución propia. Nettesheim y Quarthal, «La reforma de las Constituciones de los Länder», 282-83.

[11] Djeffal, «AI, Democracy and the Law», 279. En este mismo sentido, parece situarse el mandato de nuestra Constitución en virtud del cual el legislador debe limitar el uso de la informática para el ejercicio pleno de los derechos fundamentales. Art. 18(4) CE: *La ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos*.

[12] En este sentido: *Los sistemas de IA deberían respaldar la autonomía y la toma de decisiones de las personas, tal como prescribe el principio del respeto de la autonomía humana. Esto requiere que los sistemas de IA actúen tanto como facilitadores de una sociedad democrática, próspera y equitativa, apoyando la acción humana y promoviendo los derechos fundamentales, además de permitir la supervisión humana*. Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (HLEG-AI), *Directrices éticas para una IA fiable*, 21.)

[13] Este documento señala que: *La supervisión humana ayuda a garantizar que un sistema de IA no socave la autonomía humana o provoque otros efectos adversos. El objetivo de una IA fiable, ética y antropocéntrica solo puede alcanzarse garantizando una participación adecuada de las personas con relación a las aplicaciones de IA de riesgo elevado*. Vid. Comisión Europea, Comunicación de la Comisión «*Libro Blanco sobre la inteligencia artificial – un enfoque europeo orientado a la excelencia y la confianza*». Bruselas, 19.02.2020. COM (2020) 65 final. Disponible en: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

Resolución del Parlamento Europeo con recomendaciones destinadas a la Comisión sobre inteligencia artificial, robótica y tecnologías conexas[14].

Cómo los seres humanos pueden mantener el control sobre tecnologías cada vez más avanzadas y autónomas, manteniendo a su vez su propia autonomía, es una pregunta de enorme relevancia para el Derecho. Por ello, es necesario analizar cómo el ordenamiento vigente regula la toma de decisiones automatizada sobre la base de inferencias algorítmicas, cuestionar si dicha regulación es suficiente para abordar los retos a los que nos enfrentan tecnologías socialmente disruptivas como los sistemas de IA y discutir si las iniciativas regulatorias propuestas para renovar este marco aportan, o no, una mejora sustancial respecto de la regulación vigente. El cumplimiento de todos estos objetivos sería, desde luego, imposible de abordar en una única investigación. Por ello, es necesario acotar el alcance de este trabajo.

1. Objeto de estudio

Resulta conveniente aportar una tentativa de definición de qué se entenderá en esta investigación por "toma de decisiones automatizada basada en la elaboración de perfiles". La definición que propongo es la siguiente: «*proceso de toma de decisiones en el que, a partir del resultado de un sistema algorítmico que realiza una inferencia, clasificación o evaluación sobre una persona física, se adopta una decisión parcial o totalmente automatizada que afecta a la misma*».

Por un lado, el término "proceso" de toma de decisiones, frente a un "sistema", entiende la toma de decisiones como parte de un entorno social que precede y condiciona la

[14] Entre otros, recoge dicha resolución que el Parlamento Europeo: *(...) 10. Considera que la inteligencia artificial, la robótica y las tecnologías conexas deben adaptarse a las necesidades humanas, en consonancia con el principio según el cual su desarrollo, despliegue y uso deben estar siempre al servicio del ser humano y nunca al revés y deben tener por objeto aumentar el bienestar y la libertad individual, así como preservar la paz, prevenir los conflictos y reforzar la seguridad internacional, maximizando al mismo tiempo los beneficios ofrecidos y evitando y reduciendo los riesgos; 11. Declara que el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas de alto riesgo, en particular —pero no solo— por parte de los seres humanos, deben regirse siempre por principios éticos y estar concebidos para respetar y permitir la intervención humana y el control democrático, así como permitir la recuperación del control humano cuando sea necesario aplicando medidas de control adecuadas*. Vid. 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)). Disponible en: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_ES.html

introducción en el mismo de un sistema algorítmico[15]. Al mismo tiempo, el término "sistema algorítmico" permite abarcar distintas clases de tecnologías, más o menos punteras, y que puedan o no coincidir con lo que en distintos contextos se define como IA[16].

Además, con "sistema" se pretende hacer referencia a todo el ciclo de vida del algoritmo, incluyendo desde sus fases de diseño y desarrollo al despliegue e implementación del mismo. Esto es, el diseño y desarrollo del sistema algorítmico, a pesar de ubicarse fuera del propio proceso de toma de decisiones -puesto que se produce, por supuesto, con anterioridad a la utilización o implementación del sistema-, es un aspecto indisociable del análisis jurídico de la toma de decisiones automatizada basada en la elaboración de perfiles.

Por otro lado, hay una delegación total o parcial del proceso de toma de decisiones en una función automatizada. El "resultado" del sistema, es decir, la ejecución de la función automatizada delegada en el proceso de toma de decisiones, tiene por finalidad la evaluación, predicción o clasificación que afecte a una persona física, lo que podríamos denominar "elaboración de perfiles"[17]. Y esa delegación resulta total o parcial, lo que significa que no excluye del concepto "decisión automatizada" la posibilidad de una intervención humana mayor o menor en el proceso decisorio[18]. Al contrario, permite entender de forma amplia el proceso decisorio en el que se produce una delegación -que puede ser de distinta clase- para la ejecución automática de determinadas acciones. De esta forma, y en contra de lo que se ha entendido en gran parte de la literatura, la decisión

[15] Krupiy, «A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective», 4.

[16] En la misma línea, AlgorithmWacht entiende que este término no resulta tan difuso como el de inteligencia artificial, que invoca connotaciones de autonomía e intencionalidad de tipo humano que no deben atribuirse a los procedimientos basados en máquinas y que, además, podría excluir de su definición algunos modelos algorítmicos que -siendo considerablemente simples- no dejan de tener un impacto notable para las personas. Vid. Spielkamp (Ed.), «Automating Society. Taking Stock of Automated Decision-Making in the EU».

[17] Del mismo modo, no tomamos por el momento elaboración de perfiles como equivalente a la definición aportada en el artículo 4(4) RGPD.

[18] En este sentido, el informe de 2019 de AlgorithmWatch, en el cual justificaban por qué hablaban de toma de decisiones automatizada y no de IA, puede resultar muy clarificador: *Algorithmically controlled, automated decision-making or decision support systems are procedures in which decisions are initially—partially or completely—delegated to another person or corporate entity, who then in turn use automatically executed decision-making models to perform an action. This delegation—not of the decision itself, but of the execution—to a data-driven, algorithmically controlled system, is what needs our attention*. Spielkamp (Ed.), «Automating Society. Taking Stock of Automated Decision-Making in the EU».

automatizada no es equivalente a una automatización plena de la toma de decisiones[19] y comprende distintas posibilidades, desde el apoyo a los responsables de la toma de decisiones hasta los procesos de toma de decisiones *completamente* automatizados, en una gran variedad de contextos[20].

Por último, si la ejecución de la función automática pone el foco en la finalidad de la misma sobre una persona física (inferencia, clasificación o evaluación), el proceso de toma de decisiones pone el foco sobre el efecto que produce el proceso en la persona física con la "afectación" a la misma[21].

En cuanto al enfoque particular que esta investigación desarrolla acerca de la participación humana en el proceso de decisiones automatizada, por el momento es oportuno precisar que se explora fundamentalmente las distintas formas en las que el Derecho normativiza esta participación humana en la toma de decisiones automatizada. Como veremos, en el proceso de toma de decisiones automatizado el papel de la interacción humana puede cumplir objetivos normativos muy diversos, pudiendo variar también la relación de esta interacción con la elaboración de perfiles y con el significado que el ordenamiento jurídico otorgue a dicha inferencia algorítmica.

2. Justificación, método y estructura de la tesis

Esta investigación versa sobre la toma de decisiones automatizada basada en la elaboración de perfiles en la normativa europea sobre protección de datos. La introducción realiza una contextualización del objeto de estudio, tanto desde un punto de vista técnico, como acercando el objeto a su análisis socioeconómico y político-jurídico.

Como se verá, en todo sector público o privado pueden utilizarse sistemas algorítmicos en procesos de toma de decisiones automatizada. La bibliografía analizada al comenzar esta investigación arrojó, en lo que se refiere al acercamiento normativo a la utilización

[19] Si es que esa automatización "plena" puede existir hoy día. Sí se hará distinción jurídica posteriormente entre lo que el RGPD entiende por "decisión basada únicamente en el tratamiento automatizado" y el resto de decisiones automatizadas que son objeto de regulación en su artículo 22. Vid. Capítulo 2. Toma de decisiones automatizada en el RGPD: el artículo 22 en la unidad de cuidados intensivos. Diagnóstico y propuestas terapéuticas para su recuperación.

[20] Araujo et al., «In AI we trust? Perceptions about automated decision-making by artificial intelligence».

[21] Es decir, la toma de decisiones no se concibe desde una posición que podríamos denominar "finalista" -¿hay una intención de tomar una decisión?-, sino desde el efecto "externo" que el proceso produce sobre la persona evaluada.

de estos sistemas o modelos algorítmicos, una serie de elementos comunes sobre los que pivotan los procesos de toma de decisiones automatizada en distintos contextos.

El intento de sistematizar dichos elementos constituye el primer capítulo de esta investigación. Partiendo del análisis bibliográfico realizado, y con apoyo de las distintas iniciativas para la regulación europea de los sistemas de IA de "alto riesgo" para una amplia variedad de contextos, trato de realizar una sistematización de los elementos comunes que conforman la gobernanza de los sistemas algorítmicos utilizados en procesos de toma de decisiones automatizada[22]. Si esta sistematización resultaba necesaria para el propio desarrollo del resto de la investigación, puede entenderse también como un resultado en sí mismo que he denominado "marco teórico de la toma de decisiones automatizada basada en la elaboración de perfiles", y servir de referencia a cualquier jurista que se acerque a la materia.

A partir de aquí, resultaba indispensable analizar en profundidad la regulación de la toma de decisiones automatizada basada en la elaboración de perfiles en el Reglamento General de Protección de Datos[23]. El ámbito de aplicación de este instrumento normativo es transversal en lo que se refiere al tratamiento de datos personales que se produce en la toma de decisiones automatizada -para la fase de implementación de los sistemas algorítmicos-, y por ello, la doctrina lo ha tomado como referencia para el estudio de los sistemas algorítmicos que realizan esta clase de tratamiento–tanto en derecho público como en derecho privado–. De esta forma, el artículo 22 del Reglamento que declara el derecho a no ser objeto de decisiones individuales automatizadas, incluida la elaboración de perfiles, se convierte en protagonista ineludible de este análisis, con especial énfasis sobre los mecanismos de gobernanza basados en la intervención humana que incluye esta disposición.

El análisis de la normativa de protección de datos se divide en tres capítulos distintos.

En primer lugar, el análisis jurídico comienza por realizar un acercamiento a los orígenes de la regulación de la toma de decisiones automatizada y la elaboración de perfiles en la

[22] Las fuentes normativas que se toman como referencia para el desarrollo de este marco teórico son más heterogéneas que en el resto de la investigación. En cualquier caso, son recurrentes las referencias al RGPD para ubicar dicho marco en el posterior desarrollo de la investigación.

[23] Esta investigación es fundamentalmente un estudio sobre el Reglamento General de Protección de Datos que resulta directamente aplicable en el ordenamiento jurídico español.

normativa de protección de datos y de su relación con la protección de los derechos a la vida privada y a la protección de datos, indagando también en los precedentes y disposiciones análogas del entorno normativo europeo a la vigente disposición kafkiana[24]: el artículo 22 RGPD. A continuación, se analiza la ubicación de esta disposición en el Reglamento y, desde unas líneas más bien generales, su problemático contenido. Este capítulo pondrá de manifiesto las dos notas características que hacen de esta disposición un derecho de segunda categoría: su falta de claridad y su escasa aplicación.

En segundo lugar, se propone una interpretación que ofrezca vías útiles para la aplicación del vilipendiado artículo 22 RGPD. Partiendo de desviar el foco del análisis doctrinal sobre los derechos de información y acceso y, en particular, la discusión relativa a la existencia o no de un derecho a una explicación para las decisiones automatizadas y su extensión. A través de los tres pilares interpretativos que se proponen, pretendo poner de relieve que el conjunto de mecanismos de gobernanza, ex ante y ex post, que pone a disposición de la persona interesada el RGPD en la regulación de la toma de decisiones -con especial atención sobre la intervención humana significativa exigida por el Reglamento- es más amplia de lo que muchos análisis jurídicos han propuesto hasta la fecha.

Por último, tratando de superar las limitaciones que arroja la propuesta interpretativa del capítulo anterior -centrada en el ejercicio de los derechos individuales reconocidos por el RGPD-, se realiza un análisis de la regulación de la toma de decisiones automatizada basada en la elaboración de perfiles desde la piedra angular del RGPD: la responsabilidad sobre el cumplimiento normativo y su demostración. Este capítulo parte de la constatación de las dificultades para garantizar una intervención humana significativa desde una perspectiva individual, para reclamar una intervención humana significativa y demostrable que encuentra su acomodo en una herramienta fundamental -partiendo del principio nuclear de la responsabilidad- en la regulación de la toma de decisiones automatizada basada en la elaboración de perfiles: la evaluación de impacto de protección de datos.

[24] Sobre el origen de este sobrenombre que ha recibido esta disposición vid. Capítulo 2. Toma de decisiones automatizada en el RGPD: el artículo 22 en la unidad de cuidados intensivos. Diagnóstico y propuestas terapéuticas para su recuperación. Apartado 2. La disposición kafkiana: Artículo 22 del RGPD. Razones para su ingreso en la UCI.

Al final de cada capítulo, a modo de cierre, se recogen una serie de reflexiones provisionales. Aunque algunas de estas reflexiones servirán de apoyo para las conclusiones de esta investigación, el objetivo de estos apartados no es exponer dichas conclusiones propiamente, sino resaltar de forma telegráfica algunos aspectos clave resultado del análisis realizado en cada capítulo. La investigación finaliza con la exposición de las conclusiones que han podido ser extraídas de este trabajo destacando las vulnerabilidades halladas en la legislación vigente e identificando algunas de las posibles mejoras que podría incorporar el ordenamiento jurídico.

Sin ser esta tesis doctoral el resultado de una compilación de artículos, la presente investigación se ha desarrollado en paralelo a otros trabajos de investigación menos extensos en los que he tenido la oportunidad de abordar de forma tangencial el objeto de estudio aquí presentado. Muchos de los argumentos desarrollados en dichos trabajos me han permitido reforzar y refutar a partes iguales algunas de las hipótesis que se plantean en esta investigación. Aunque estos trabajos se citan posteriormente, me gustaría recopilarlos en este apartado y expresar también mi agradecimiento a las personas que me han acompañado en dichas investigaciones, especialmente en forma de coautoría, pero también a quienes desempeñaron tareas de revisión (anónima o no), edición, etc., y que contribuyeron en menor o mayor medida a los resultados de estos trabajos.

En orden cronológico de publicación o pre-publicación:

- En coautoría con Carlos María Romeo Casabona, «Inteligencia artificial aplicada a la salud : ¿qué marco jurídico?» para la Revista de Derecho y Genoma Humano: Genética, Biotecnología y Medicina Avanzada.
- Como autor único, « Análisis jurídico de la toma de decisiones algorítmica en la asistencia sanitaria» en *La regulación de los algoritmos*.
- En coautoría con Iñigo de Miguel Beriain, «Big Data Analysis y Machine Learning en medicina intensiva: identificando nuevos retos ético-jurídicos», para Medicina Intensiva.
- En coautoría con José Antonio Castillo Parrilla, «Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: el caso SyRI», para la Revista Chilena de Derecho y Tecnología.
- En coautoría con Carlos María Romeo Casabona, Pilar Nicolás Jiménez e Iñigo de Miguel Beriain, «Proyecte IA RXCOVID 19: Programa Intelligència Artificial al SISCAT», para Fundació TICSalut del Departament de Salut de la Generalitat de Catalunya.

- En coautoría con Iñigo de Miguel Beriain y Begoña Sanz, «Machine learning in the EU health care context: exploring the ethical, legal and social issues», para Information, Communication & Society.
- En coautoría con Iñigo de Miguel Beriain, «Inteligencia artificial, personas mayores y biomedicina la vulnerabilidad en el debate ético-jurídico» en *Soluciones tecnológicas para los problemas ligados al envejecimiento: cuestiones éticas y jurídicas*.
- Como autor único, «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» para IUS ET SCIENTIA.
- Como autor único, «Modelos algorítmicos, sesgos y discriminación» en *FODERTICS 9.0: Estudios sobre tecnologías disruptivas y justicia*.
- En coautoría con Aritz Obregón, «La supervisión humana de los sistemas de inteligencia artificial de alto riesgo. Aportaciones desde el Derecho Internacional Humanitario y el Derecho de la Unión Europea», para la Revista Electrónica de Estudios Internacionales.
- Como autor único, «Automated decision-making under Amsterdam's District Court judgements: Drivers v. Uber and Ola», en *Time to reshape the digital society. 40th anniversary of the CRIDS*.
- En coautoría con Andrea Perin, «Lección 20. Inteligencia artificial en el ámbito sanitario» en *Manual de Bioderecho: (adaptado a la docencia en ciencias, ciencias de la salud y ciencias sociales y jurídicas)*.
- En coautoría con Paul de Hert, «When GDPR-Principles Blind Each Other: Accountability, Not Transparency, at the Heart of Algorithmic Governance», para European Data Protection Law Review.
- En coautoría con Iñigo de Miguel Beriain, Pilar Nicolás Jiménez, Maria Jose Rementeria, Davide Cirillo, Atia Cortés y Diego Saby, «Auditing the quality of datasets used in algorithmic decision-making systems» para el Panel para el futuro de la ciencia y la tecnología (STOA) del Parlamento Europeo.
- En coautória con Paul de Hert, «Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities» (pre-publicación).

Salvo indicación en contrario, en esta investigación la traducción de las citas originales en lengua inglesa al castellano ha sido realizada por mí. En algunos casos he considerado apropiado mantener la cita original.

INTRODUCCIÓN A LA GOBERNANZA Y SUPERVISIÓN HUMANA DE LA TOMA DE DECISIONES AUTOMATIZADA BASADA EN LA ELABORACIÓN DE PERFILES

INTRODUCCIÓN A LA GOBERNANZA Y SUPERVISIÓN HUMANA DE LA TOMA DE DECISIONES AUTOMATIZADA BASADA EN LA ELABORACIÓN DE PERFILES

Para poder abordar el objeto de estudio de esta investigación es indispensable realizar, en primer lugar, una adecuada contextualización que permita fijar el entorno en el que se desarrolla y pervive dicho objeto. Y para ello, a su vez, será necesario analizar una doble perspectiva relativa a este fenómeno: técnica y socioeconómica.

Por un lado, es obligado referirnos a los modelos algorítmicos predictivos que sirven para dicha elaboración de perfiles y toma de decisiones automatizada. El desarrollo científico y tecnológico ha permitido el éxito de esta clase de modelos para la resolución de muy diversos problemas, lo cual ha sido un factor determinante en la extensión de su uso y despliegue en muy diversos ámbitos. El éxito de estos modelos ha sido tal, que ha traído de vuelta al debate público el término "inteligencia artificial" para referirse a la clase de tecnologías que se utilizan para el desarrollo de estos modelos.

Por otro lado, es imposible entender el "éxito" de una determinada tecnología sin comprender el contexto social en el que se desarrolla esta "revolución" y los condicionantes que la hacen plausible. La superación de una concepción neutral de lo tecnológico es también una de las premisas de esta investigación. Los cauces por los que discurre la revolución tecnológica vienen en cualquier caso determinados por diversos marcos socioeconómicos, también de ideologías políticas, que no pueden ser obviados a la hora de realizar una investigación jurídica sobre la materia. En última instancia, me ocuparé del papel que el Derecho ha ocupado en este fenómeno.

En definitiva, la presente contextualización pretende alumbrar distintos puntos de vista desde los cuáles puede comprenderse la presente "revolución" tecnológica, tras la que la toma de decisiones automatizada basada en la elaboración de perfiles adquiere una relevancia jurídica sin precedentes en el momento actual. Tras finalizar esta contextualización, esta introducción finaliza planteando en forma de preguntas, y sin ánimo de exhaustividad sino como mero aperitivo, algunas de las cuestiones que se resolverán en esta investigación.

1. Contextualización técnica: Modelos algorítmicos y sistemas de inteligencia artificial

Algoritmo -sustantivo-: «*Palabra utilizada por programadores cuando no quieren explicar lo que han hecho*»[25].

La popularización de algunos términos como inteligencia artificial, algoritmos o *big data* ha generado desinformación y confusión sobre dichos conceptos y sus distintas prácticas [26]. Incluso en ámbitos estrictamente académicos, donde esta confusión ha contribuido a exagerar las capacidades y promesas de las tecnologías englobadas por estos términos[27]. En este apartado, recogiendo fuentes autorizadas en la materia, trato de exponer las distintas tecnologías que serán objeto de estudio en esta investigación y su funcionamiento más básico. Lo haré partiendo de la noción de "inteligencia artificial".

Derivar la toma de decisiones a máquinas se asocia inevitablemente con que dichos artefactos contengan alguna forma de inteligencia. Desde luego, sería harto cuestionable confiar las decisiones tradicionalmente tomadas por seres humanos a entidades que no considerásemos capaces de una mínima inteligencia, o al menos ello no dejaría en muy buen lugar a los humanos. Ahora bien, definir qué entendemos por inteligencia artificial (IA en adelante) en relación con nuestro objeto de estudio, o al menos encontrar una definición consensuada, no resulta sencillo[28].

---

[25] Me ha sido imposible encontrar el origen de esta broma habitualmente utilizada en conferencias y publicaciones sobre la materia. Tomemos por referencia la definición de Yeung: *En su sentido más amplio, los algoritmos son procedimientos codificados para resolver un problema transformando los datos de entrada en un resultado deseado*. Aunque como veremos después el término "algoritmo" es utilizado para referirse a modelos de aprendizaje automático o sistemas de IA en general, sobre las distintas modulaciones que recibe lo percibido como "algorítmico" vid. Yeung, «Algorithmic regulation: A critical interrogation». Yeung 2018.

[26] McQuillan, «Data Science as Machinic Neoplatonism», 254. Para ver hasta qué punto son maleables estos términos e intercambiables entre sí en la actualidad, vid. Katz, «Manufacturing an Artificial Intelligence Revolution».

[27] Hagendorff y Wezel, «15 challenges for AI: or what AI (currently) can't do», 355.

[28] Para hacernos una idea de lo complejo – y de la importancia – de esta tarea, el Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (AI HLEG, en adelante) consideró necesario publicar un documento anexo a sus Directrices Éticas para una IA Fiable para aclarar ciertos aspectos de la IA como disciplina científica y como tecnología en relación con la definición adoptada. A definition of AI: Main capabilities and scientific disciplines (April 2019): https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

La entrada sobre este término de la Enciclopedia de la Filosofía de Stanford dice que definir con precisión una disciplina, a satisfacción de todas las partes interesadas en la misma, es un gran desafío -tal vez irrealizable-[29]. Y es que, si hablamos de partes interesadas, las disciplinas que confluyen en el desarrollo de la IA son muy diversas, entre otras, las ciencias de la computación, la lingüística, las matemáticas, la estadística, la biología, la neurociencia o la psicología; lo cual dificulta aún más la tarea de encontrar esa definición de consenso.

A esta complicación, se suma una concepción antropocentrista[30] de la IA que es, además, de doble sentido; por un lado, la concepción de inteligencia suele definirse desde lo humano[31] y, por otro, las preguntas sobre la IA llevan inevitablemente a preguntarse qué significa ser humano y dónde están exactamente los límites de lo que nos define[32]. No parece apropiado restringir la noción de "inteligencia" a lo que requeriría la inteligencia si fuera hecha por humanos[33]. La segunda cuestión, que nos llevará a cuestionar cómo se adoptan y regulan las decisiones humanas, sí será más explorada en esta investigación[34].

Dado el alto grado de aceptación que ha tenido este documento, podemos adoptar la definición del Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (AI HLEG, en adelante): «*Los sistemas de inteligencia artificial (IA) son 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*

[29] Bringsjord y Govindarajulu, «Artificial Intelligence».

[30] Martínez Zorrilla propone una definición de IA fuera de sesgos antropocéntricos, entendiendo la IA como la capacidad de procesar la información obtenida, ya sea por medios propios o proporcionada por terceros, para producir de modo autónomo un resultado como consecuencia de dicho procesamiento. Considera dicha capacidad de forma gradual – de más básico a más sofisticado – y aplicable tanto a animales como a sistemas informáticos. Seminario impartido el 8 de febrero de 2019 en la Facultad de Derecho de la Universidad de Girona. "La decisión judicial automatizada: entre la ciencia y la ficción". Disponible en: https://www.youtube.com/watch?v=e-WI3U1_BI0

[31] Como ejemplo, Kurzweil propone definir la IA como el intento de construir máquinas que realizan funciones que requieren inteligencia cuando son realizadas por personas. Vid. Kurzweil, *The age of inteligent machines*.

[32] Musa Giuliano, «Echoes of myth and magic in the language of Artificial Intelligence», 2.

[33] Müller, «Ethics of Artificial Intelligence and Robotics».

[34] Especialmente cuando sean exploradas las razones del ordenamiento jurídico para mantener a seres humanos en los procesos decisorios automatizados.

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. Los sistemas de IA pueden utilizar normas simbólicas o aprender un modelo numérico; también pueden adaptar su conducta mediante el análisis del modo en que el entorno se ve afectado por sus acciones anteriores»[35]. A lo que añade que la IA, como disciplina científica, incluye varios acercamientos y técnicas como el aprendizaje automático, el razonamiento automático o la robótica; campos a los que nos referiremos más adelante.

Esta definición va en la misma línea que otras propuestas, especialmente a la hora de abordar la regulación expresa de estos sistemas. Me refiero a que se trata de definiciones deliberadamente amplias que abarcan incluso sistemas que no son necesariamente complejos[36], y que desde luego incluyen a los modelos "de actualidad" basados en aprendizaje automático a los que me referiré más adelante. Dicha amplitud tiene una doble lectura, primera, que la IA engloba una multitud de técnicas muy diferentes entre sí a las que me referiré al recoger brevemente el recorrido histórico de la misma. Segundo, que el propio concepto de IA tiene una fuerte carga simbólica que trasciende de la definición que puede adoptarse a partir de la consideración de determinadas técnicas o desarrollos tecnológicos.

1.1. Breve recorrido histórico por la inteligencia artificial: una batalla entre dos formas contrapuestas de entenderla

[35] Comisión Europea, Libro Blanco sobre la inteligencia artificial, 48. La definición propuesta por la Comisión en la Propuesta de Reglamento por el que se establecen normas armonizadas en materia de inteligencia artificial (en adelante propuesta de Reglamento AIA o AIA) limita la "actuación" del sistema a la generación de información de salida como "contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa" en su artículo 3(1) AIA. Vid. Comisión Europea, Propuesta de Reglamento por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial). Bruselas, 21.4.2021. COM(2021) 206 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex:52021PC0206

[36] La amplitud de la definición de la inteligencia artificial es, en realidad, una cuestión jurídica de primera magnitud. En el seno del procedimiento legislativo ordinario de la propuesta de Reglamento AIA, se debate si deben o no acotarse los sistemas y modelos algorítmicos incluidos por la Comisión en la versión inicial como "inteligencia artificial". Sin entrar en las particularidades de esta discusión, el debate gira en torno a si la regulación debe afectar exclusivamente a los sistemas más complejos. Esta posición despierta el temor de dejar sin regulación sistemas que afectan a personas físicas y ponen en alto riesgo sus derechos fundamentales. Este temor pone de manifiesto que los efectos de la aplicación de estas tecnologías no dependen necesariamente de la utilización de las tecnologías más avanzadas o inteligentes. Vid. Bryson, «Europe Is in Danger of Using the Wrong Definition of AI». Disponible en: https://www.wired.com/story/artificial-intelligence-regulation-european-union/

El término "inteligencia artificial" fue acuñado en el verano de 1956, en la Conferencia de Dartmouth (Dartmouth Summer Research Project on Artificial Intelligence) organizada por John McCarthy y celebrada en la universidad Dartmouth College, ubicada en Hanover, New Hampshire. Sin embargo, en 1950 Alan Turing ya propuso considerar la cuestión "¿pueden pensar las máquinas?" en una publicación en la revista Mind[37] . En términos generales podemos pensar la IA como un campo científico en el que se pretende que los sistemas informáticos desarrollen procesos lógicos semejantes a la mente humana, es decir, que tiene por objetivo explicar y emular el comportamiento inteligente en términos de procesos computacionales[38]. Y a este objetivo es posible llegar por distintas vías. Esta posibilidad se hizo patente de forma temprana en la historia de la IA, cuyo desarrollo suele exponerse a partir de dos olas diferenciadas[39].

Por un lado, y en un primer momento, surgió el desarrollo de la IA simbólica, ahora también conocida como "*good old-fashioned AI*". Estos sistemas utilizan símbolos matemáticos para representar objetos y la relación entre ellos, y fue el enfoque dominante hasta la década de los años noventa. Los conocidos como sistemas expertos o basados en reglas son los que más han perdurado en el tiempo. Para el funcionamiento de estos sistemas, una persona experta en el ámbito de la aplicación crea reglas precisas que un ordenador puede seguir, paso a paso, para decidir cómo responder de forma inteligente a una situación determinada -habitualmente siguiendo instrucciones lógicas en formato "Si-Entonces"-[40]. Su razonamiento se considera "fuerte" dado que refleja la lógica y reglas de quienes sus programadores humanos, no obstante, al igual que ocurrió con el resto de sistemas basados en la IA simbólica, en la década de 1980 quedó claro que estos sistemas no eran capaces de ofrecer un rendimiento adaptado a la fluidez de los símbolos, los conceptos y el razonamiento en la vida real[41].

La segunda ola abarca los sistemas de aprendizaje automático basados en la creación de modelos predictivos a partir de los datos. Estos enfoques se inspiran más en la estadística

[37] Turing, «Computing Machinery and Intellingence».

[38] del Río Solá, López Santos, y Vaquero Puerta, «La inteligencia artificial en el ámbito médico», 113.

[39] Muy recomendable el libro de Margaret Boden en el que se desarrolla ampliamente la discusión entre los distintos campos. Vid. Boden, *Inteligencia Artificial*.

[40] Boucher, «Artificial intelligence: How does it work, why does it matter, and what can we do about it?», 2.

[41] Waldrop, «News Feature: What are the limits of deep learning?», 1074.

que en la neurociencia o la psicología y estaban orientados a realizar tareas específicas más que a captar la inteligencia general[42]. Aunque los conceptos en los que se basan estos enfoques son tan antiguos como la IA simbólica, no se aplicaron de forma generalizada hasta después del cambio de siglo, cuando inspiraron el actual resurgimiento de la IA[43]. De hecho, uno de los métodos de mayor éxito en la actualidad sobre el que entraremos más adelante, las redes neuronales de aprendizaje profundo, fueron ideadas en la década de los 80, pero la capacidad computacional por entonces disponible hacía de estos modelos soluciones muy poco útiles.

En resumen, lo que caracteriza a estos modelos de aprendizaje automático es el aprendizaje a partir de datos, frente a la IA simbólica en la que el razonamiento se producía a partir de reglas prefijadas. El aprendizaje automático revierte este proceso en cierto sentido, el "aprendizaje" aquí describe el proceso de la búsqueda automática de mejores y más útiles predicciones para los datos con los que alimentamos al modelo, significa que la máquina puede mejorar en su tarea programada, rutinaria y automatizada a partir de ejemplos que se le facilitan[44].

Los definidos como inviernos de la IA tienen una importancia capital en el devenir histórico del desarrollo tecnológico de estas tecnologías, más si cabe cuando incorporamos esta perspectiva histórica al análisis actual. La literatura señala habitualmente que dichos periodos invernales se produjeron a finales de los años 60 y 80 respectivamente -vid. ilustración-. Los inviernos de la IA se caracterizan por ser periodos de decepción, pérdida de confianza y reducción de la financiación; ello no quiere decir que la innovación y el desarrollo de sistemas se paralizase por completo, pero sí que, como poco, se rehuía conscientemente la etiqueta "inteligencia artificial"[45].

El *hype* o bombo publicitario es un antecedente común a estos periodos, dicho *hype* llevó a inversiones desmedidas y publicidad exagerada que tuvieron una relación directa con el

[42] Mitchell, «Why AI is Harder than We Think».

[43] Boucher, «Artificial intelligence: How does it work, why does it matter, and what can we do about it?», 3.

[44] Pero no que la máquina adquiera conocimiento, sabiduría o agencia, a pesar de lo que el término aprendizaje pueda implicar, vid. Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, 237.

[45] Mitchell, «Why AI is Harder than We Think».

posterior hundimiento de la inversión en investigación para el desarrollo de la IA[46]. En esta línea, no son pocas las voces que han manifestado su preocupación por la posible aparición de un nuevo invierno de la IA[47]. Una vez más, las predicciones sobre el futuro no son objeto de esta investigación, pero lo que sí parece evidente es que nos hallamos nuevamente en un momento histórico de *hype* de la IA.

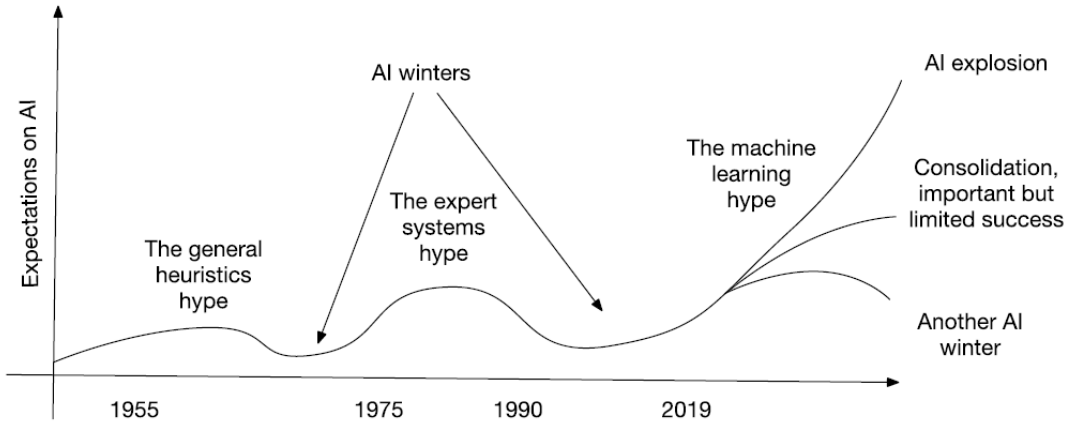

Ilustración 1. Sartor y Lagioia 2020, 5

1.2. El presente de la inteligencia artificial: modelos algorítmicos de aprendizaje automático basados en el tratamiento masivo de datos

En el apartado anterior adelantaba que los modelos de aprendizaje automático que forman parte de esa segunda ola de la IA, son protagonistas del resurgimiento actual de este campo. Dado que esta perspectiva desde la que desarrollar la IA la encontramos ya en las primeras discusiones en la disciplina, hemos de buscar las razones de su éxito actual en otros factores.

[46] Bentley et al., «¿Debemos temer a la inteligencia artificial? Análisis en profundidad», 13.

[47] En sus propias palabras, Gary Marcus: *When a high-profile figure like Andrew Ng writes in the Harvard Business Review promising a degree of imminent automation that is out of step with reality, there is fresh risk for seriously dashed expectations. Machines cannot in fact do many things that ordinary humans can do in a second, ranging from reliably comprehending the world to understanding sentences. No healthy human being would ever mistake a turtle for a rifle or parking sign for a refrigerator*. Vid. Marcus, «Deep Learning: A Critical Appraisal». También Bentley et al. sobre los posibles efectos devastadores del *hype*: *No se crean el bombo publicitario. Somos terribles prediciendo el futuro y, casi sin excepción, las previsiones (incluso las realizadas por expertos mundiales) son completamente erróneas. (…) Las grandes afirmaciones llevan a una gran publicidad, que conduce a una gran inversión y nuevas normativas. Y entonces es cuando sacude la realid ad inevitable. La IA no está a la altura del bombo publicitario. La inversión se seca. La normativa ahoga la innovación. Y la IA se convierte en una expresión sucia que nadie osa pronunciar. Otro ocaso de la IA destruye el progreso*. Bentley et al., «¿Debemos temer a la inteligencia artificial? Análisis en profundidad», 13.

Por un lado, aparece como factor esencial el aumento masivo de la disponibilidad de datos[48], la era del *big data* o de los macrodatos permite la generación y recogida de cantidades masivas de datos que abarcan nuestra cotidianidad en su sentido más amplio. Al ser modelos con capacidad para manejar un número de variables inabordable por la programación clásica[49], la disponibilidad de los macrodatos es propicia para su éxito[50].

Por otro lado, el aumento exponencial de la capacidad computacional ha permitido que estos modelos de la segunda ola de la IA resulten de utilidad, junto con el hecho de que dicho desarrollo se haya acompañado de un abaratamiento de las herramientas que permiten la recolección y almacenaje de los datos y su procesamiento[51].

Si el aprendizaje computacional es un campo de otros tantos dentro de la IA, el propio campo del aprendizaje computacional abarca un extenso abanico de posibilidades algorítmicas. Con carácter general, la idiosincrasia del aprendizaje automático reside en la capacidad de inferir reglas de forma autónoma a partir de los patrones que se observan en las bases de datos que les son proporcionadas -datos de entrenamiento-, mejorando su rendimiento a partir de esos ejemplos. Es decir, el "aprendizaje" consiste en la capacidad de la máquina para mejorar en su tarea programada, rutinaria y automatizada[52], y podemos distinguir a su vez entre distintas clases de aprendizaje[53]. Los modelos de aprendizaje automático son en lo fundamental modelos estadísticos, esos patrones que el algoritmo

[48] Boucher, «Artificial intelligence: How does it work, why does it matter, and what can we do about it?», 4.

[49] Obermeyer y Emanuel, «Predicting the Future — Big Data, Machine Learning, and Clinical Medicine», 1216.

[50] Es decir, los modelos de aprendizaje automático, como sistemas estadísticos que utilizan los datos para mejorar su rendimiento para tareas determinadas, son el enfoque elegido para generar valor a partir del "agotamiento de datos" de las actividades humanas digitalizadas. Veale, Van Kleek, y Binns, «Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making», 3.

[51] Kamarinou, Millard, y Singh, «Machine Learning with Personal Data», 4.

[52] Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, 89.

[53] Con carácter general: supervisado, no supervisado y por refuerzo. El aprendizaje supervisado implica que los datos de entrenamiento se etiquetan de antemano, indicando a los algoritmos cómo deben calificarse (clasificación). El aprendizaje no supervisado implica que los datos de entrenamiento se examinan en busca de patrones, para detectar clases de datos potencialmente relevantes (agrupación), lo que también permite al algoritmo detectar anomalías que pueden ser ignoradas en el aprendizaje supervisado. El aprendizaje por refuerzo implica que los datos de salida se clasifican como correctos (invocando una recompensa) o incorrectos (invocando un castigo), constituye en cierto sentido la versión mecanizada del conductismo. Para una explicación más extensa vid. Boden, *Inteligencia Artificial*.

extrae de los datos representan correlaciones estadísticas cuyo hallazgo escapa por lo general a la capacidad humana para detectar patrones de forma sistematizada[54].

A su vez, hay distintas técnicas de clasificación cuya complejidad varía; por ejemplo, los árboles de decisión se consideran habitualmente como los modelos predictivos más "simples", mientras que las redes neuronales artificiales, que son modelos matemáticos inspirados en el comportamiento biológico de las neuronas, son más "complejas"[55].

Esta clase de modelos algorítmicos son ya parte cotidiana de nuestras vidas, son múltiples los mecanismos de clasificación y evaluación con impacto social basados en el aprendizaje automático: filtros de *spam*, detección de fraude con tarjetas de crédito, motores de búsqueda, tendencias de actualidad, publicidad personalizada, calificación de seguros o préstamos y un largo etcétera de aplicaciones[56]. Incluso sin ser conscientes de ello, millones de decisiones por segundo que afectan a nuestras vidas son tomadas por sistemas basados en el aprendizaje automático y la tendencia es a aumentar esa delegación de tareas en los mismos[57].

No obstante, el aprendizaje de estos modelos no deja de tener sus limitaciones y en esta investigación trato de abordar la dimensión jurídica de las mismas. Aunque se profundizará sobre muchos más aspectos, sí merece la pena mencionar que habitualmente

[54] Acerca de la similitud que guardan el funcionamiento de los modelos de aprendizaje automático y la estadística y, sobre todo, de la similitud que guardan las limitaciones que se presentan a la hora de tratar de resolver un problema aplicando o bien un modelo algorítmico, o bien estadística en su sentido más clásico, no puedo dejar de recomendar el artículo de Malik. Malik, «A Hierarchy of Limitations in Machine Learning».

[55] El aprendizaje profundo o *deep learning* basado en redes neuronales artificiales, que se caracteriza por componerse de varias capas de entrada, salida y ocultas, es el ejemplo prototípico al que suele acudirse a la hora de hablar de modelos complejos, especialmente de la caja negra a la que se hace referencia más abajo. En este tipo de aprendizaje las mejores representaciones para los datos que el modelo va aprendiendo se realizan en diferentes capas, esto es, se reconocen los patrones en los datos de entrada en diferentes niveles jerárquicos; así, los datos de entrada fluyen desde la capa de entrada a la capa de salida. De forma simplificada, podríamos decir que cada capa está formada por unidades e interconectadas configuran una red neuronal y, al igual que las sinapsis cerebrales, esas conexiones entre unidades son o excitadoras o inhibitorias y varían de peso o fuerza. En la red neuronal el aprendizaje se produce por medio de la retropropagación, el algoritmo compara el resultado obtenido en la capa de salida con el resultado deseado y asume que el error en la unidad de salida se debe a errores en las unidades conectadas con ella, con lo que para corregirse realiza ajustes en los pesos asignados en la red desde la capa de salida a la capa de entrada –hacia atrás, de ahí el término retropropagación–. Boden, *Inteligencia Artificial*, 84 y ss.

[56] Entre otros, el informe sobre IA de la Agencia de los Derechos Fundamentales (FRA) ofrece un catálogo sobre los usos de esta clase de sistemas tanto en el sector privado como por la administración pública. Vid. Agencia de los Derechos Fundamentales de la Unión Europea (FRA), «Getting the future right - Artificial intelligence and fundamental rights».

[57] Carabantes, «Black-box artificial intelligence: an epistemological and critical analysis», 2.

la literatura se refiere a estos modelos como "cajas negras" por la opacidad con la que operan -veremos el contenido polisémico de este término-, que entorpecen la capacidad para comprender por parte de los operadores humanos el funcionamiento del sistema entendido en sentido amplio; es decir, tanto los patrones correlativos que el modelo establece en su entrenamiento, como el peso concreto que dichos patrones adquieren a la hora de adoptar una decisión en particular. Esta particularidad de las cajas negras ha traído discusiones -que se han reflejado también en el ámbito jurídico- acerca de la pertinencia de escoger modelos de aprendizaje automático en función de su capacidad predictiva o de su interpretabilidad o explicabilidad[58].

En definitiva, la conjunción de los factores aquí expuestos nos lleva a hablar de una nueva forma de producción de conocimiento[59]. Quizás esto último es lo más trascendente de todo -al menos desde la óptica de un jurista- cuando analizamos los modelos algorítmicos de aprendizaje automático basados en el tratamiento masivo de datos, la decisiva transformación de las originales tecnologías de la información en tecnologías de producción de conocimiento, puesto que el objetivo del aprendizaje automático en el tratamiento de datos no es la mera comunicación de los datos, sino aprender de ellos y, al hacerlo, crear nuevos conocimientos[60].

1.3. Inteligencia artificial general en el imaginario colectivo: alquimia y charlatanes

Cada vez que se hace referencia a la inteligencia artificial, se produce una brecha entre lo que imaginamos y lo que es, realmente y a día de hoy, la inteligencia artificial.

A un lado de esa brecha, se sitúa Samantha, el sistema de IA de la película *Her*, capaz no solo de consolar al protagonista, Theodore, en el duelo emocional que sufre tras la ruptura

[58] La premisa de esta discusión asume que los modelos más complejos son por sí más precisos que los modelos más sencillos y cuyas correlaciones son interpretables por los humanos, lo cual no es necesariamente cierto para todos los contextos de aplicación de los modelos. En cualquier caso, es una discusión fructífera a nivel teórico y no puedo dejar de recomendar el artículo de Iñigo de Miguel y Antonio Diéguez sobre la misma y el papel de los modelos predictivos en la innovación científica. de Miguel y Dieguez, «¿Explicar o predecir?» Disponible en: https://www.investigacionyciencia.es/revistas/investigacion-y-ciencia/al-rescate-del-coral-837/explicar-o-predecir-20016

[59] Yeung, «Algorithmic regulation: A critical interrogation», 506.

[60] Gellert, «Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms?»

de su anterior relación sentimental, sino también de enamorarse mutuamente[61]. Al otro lado, Siri, junto a otros asistentes inteligentes de distintos sistemas operativos, es capaz de recomendar restaurantes cercanos y adecuados al gusto de cada cual, pero está lejos de encontrar el amor, aunque ha sido hábilmente programado para responder «buscas el amor en el sitio equivocado» a aquellos usuarios que le plantean la pregunta «¿me quieres, Siri? ».

En otras palabras, esa IA que deseamos e imaginamos sería la Inteligencia Artificial General (IAG), mientras que la IA de la que disponemos es mucho menos intrigante, pero funciona sorprendentemente bien para llevar a cabo una amplia variedad de tareas[62] . Este imaginario colectivo que se representa habitualmente a través de la ciencia ficción, crea unas narrativas que influyen de forma decisiva en el debate público sobre la IA, influyendo en investigadores y en el conjunto de la sociedad, desplazando los pesos de los diferentes escenarios de nuestro espacio de probabilidad colectiva[63].

No, no estamos ni cerca de construir una máquina con las capacidades de un ser humano o que actúe racionalmente en todos los escenarios, sin embargo, algoritmos que tienen sus orígenes en la investigación de la IA están siendo utilizados en la actualidad para resolver un sinfín de tareas en una amplia variedad de áreas[64].

La realización de estas tareas tampoco implica que la IA comprenda esas tareas que lleva a cabo. Tal y como argumenta Mitchell en relación con el procesamiento del lenguaje natural, a menudo es difícil determinar, a partir de su rendimiento en un reto determinado, si los sistemas de IA comprenden realmente el lenguaje (u otros datos) que procesan. Ahora sabemos que las redes neuronales suelen utilizar atajos estadísticos -en lugar de demostrar realmente una comprensión similar a la humana- para obtener un alto rendimiento. En su opinión, la comprensión del lenguaje requiere la comprensión del mundo, y una máquina expuesta sólo al lenguaje no puede obtener dicha comprensión[65].

---

[61] También se expresa en esta película el temor por la singularidad de la IA, es decir, por el momento en que estos sistemas sean capaces de mejorar por sí mismos de forma que se produzca una explosión de inteligencia que quede fuera de todo control humano, y donde dicha inteligencia sea muy superior a la capacidad intelectual humana. Sobre la singularidad, vid. Boden, *Inteligencia Artificial*.

[62] Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, 10.

[63] Musa Giuliano, «Echoes of myth and magic in the language of Artificial Intelligence».

[64] Bringsjord y Govindarajulu, «Artificial Intelligence».

[65] Y ilustra este argumento con el siguiente ejemplo: *Considere lo que significa entender "El coche deportivo pasó al camión del correo porque iba más lento". Hay que saber qué son los coches deportivos*

En un trabajo más amplio, Mitchell explica que una de las falacias más comunes en este campo es el denominado "*wishful mnemonics*", que consiste en atribuir -y denominar- a una determinada función automatizada una capacidad humana, como es el caso de "aprendizaje"[66].

En el campo de la IA es habitual encontrar el recurso al símil de la alquimia: la IA es hoy lo que la alquimia fue durante siglos hasta la aparición de la ciencia moderna. A pesar de que la práctica de la alquimia estuvo rodeada de charlatanes y chamanes[67], fue una de las precursoras de las ciencias modernas. Hubert L. Dreyfus recurrió a este símil por primera vez en su ensayo "*Alchemy and Artificial Intelligence*" de 1965 que criticaba las metodologías seguidas hasta el momento en el campo de la IA[68]. En 1977, Winograd acudió de nuevo a este símil para explicar que en aquel momento la IA se encontraba en la fase de mezclar diferentes sustancias y ver qué ocurría, sin haber desarrollado aún teorías satisfactorias: «*but...it was the practical experience and curiosity of the alchemists which provided the wealth of data from which a scientific theory of chemistry could be developed*»[69]. Cuatro décadas después, Eric Horvitz, director de Microsoft Research, coincide en que en este momento lo que se hace en el campo de la IA no es ciencia, sino una especie de alquimia[70]. Ello no quiere decir que los avances en las últimas décadas no hayan sido significativos -lo han sido, no olvidemos que el paso de la alquimia a la ciencia moderna tardó siglos-, pero este símil es quizás un recurso útil para evitar la charlatanería y centrar los recursos en la búsqueda de la ciencia.

---

y los camiones del correo, que los coches pueden "adelantarse" unos a otros y, a un nivel aún más básico, que los vehículos son objetos que existen e interactúan en el mundo, conducidos por seres humanos. Mitchell, «What Does It Mean for AI to Understand?». Disponible en: https://www.quantamagazine.org/what-does-it-mean-for-ai-to-understand-20211216/

[66] Mitchell, «Why AI is Harder than We Think», 3.

[67] Permitidme añadir aquí que esta parte de la analogía es también calcada a nuestra época.

[68] Aunque esta crítica ha sido tildada en ocasiones como dura o feroz, quizás por el hecho de que Dreyfus incluyese nombres propios en la misma, sus conclusiones no son pesimistas en torno al futuro de la IA. Basado en la experiencia de la alquimia propone que deje de invertirse dinero y recursos en replicar la inteligencia humana y que se utilice para potenciar la simbiosis entre los seres humanos y las máquinas. Dreyfus, *No Alchemy and Artificial Intelligence*, 82-86..

[69] Winograd, «On some contested suppositions of generative linguistics about the scientific study of language: A response to Dresher and Hornstein's on some supposed contributions of artificial intelligence to the scientific study of language».

[70] Metz, «A new way for machines to see, taking shape in Toronto». Disponible en: https://www.nytimes.com/2017/11/28/technology/artificial-intelligence-research-toronto.html

.

En definitiva, incluso al realizar un acercamiento técnico a la materia, vemos que los términos utilizados no están exentos de una fuerza cultural y social determinante que no puede dejarse de lado en este análisis.

2. Contexto social, económico y político.

La tecnología se desarrolla inevitablemente en un contexto social, económico y político determinado. La relación entre la tecnología y dicho contexto es en todo caso bidireccional, el desarrollo técnico moldea la sociedad, y viceversa. Dicho en otras palabras, no hay forma de entender una tecnología desde una perspectiva exclusivamente técnica. En el anterior apartado veíamos brevemente cómo el imaginario colectivo tiene una influencia decisiva sobre la inversión y recursos que se destinan al desarrollo de determinadas tecnologías; a la inversa, estas tecnologías pueden jugar un papel determinante en las crisis financieras[71] o modular la identidad digital de distintos colectivos reforzando estereotipos racistas y sexistas[72].

En la reseña escrita por Rowson sobre el libro Deep Thinking de Garry Kasparov, el mítico ajedrecista que se enfrentó a la computadora *Deep Blue* -encarnando una de las historias humano vs. máquina más conocidas-, expresa refiriéndose a la IA que ésta no viene sola; que no es solo que coexista con la biología sintética, la robótica, la realidad virtual, la impresión en 3D, etc., sino que la IA surge también en un mundo con graves limitaciones ecológicas, inestabilidad económica persistente y tensiones democráticas, y debemos relacionarnos con ella en ese contexto[73].

Lo cierto es que la presente "primavera" de las tecnologías amparadas bajo el término IA en los países occidentales coincide con un momento en el que nos enfrentamos a unos retos sociales excepcionales, y en el que un pequeño puñado de empresas tecnológicas privadas ocupan una posición dominante en muchos ámbitos de la vida[74]. En este

[71] Vid. más ampliamente O'neil, *Armas de destrucción matemática*.

[72] Vid. también ampliamente Noble, *Algorithms of Oppression. How Search Engines Reinforce Racism*.

[73] Rowson, «Review of "Deep Thinking"», 49.

[74] Cowls, «'AI for Social Good': Whose Good and Who's Good? Introduction to the Special Issue on Artificial Intelligence for Social Good».Entre los múltiples roles que juegan simultáneamente las compañías tecnológicas, quizás uno de los más relevantes en este momento es que se han convertido en un actor importante en la investigación científica, como financiadores directos o indirectos y con un rol destacado en la comunidad universitaria. Prainsack, «The political economy of digital data: introduction to the special issue», 2.

momento, el término "inteligencia artificial" ha operado de forma útil como recurso retórico para los políticos que buscan establecer sus credenciales en la formulación de políticas del siglo XXI, así como para las nuevas empresas que buscan financiación inicial, sea pública o privada[75]. En el caso europeo, el impulso político por parte de la Comisión -tanto de financiación pública como de transformación normativa- responde inicialmente a la *feroz* competencia que llega desde China y EEUU, que deja a Europa muy atrasada en inversión y recursos, considerando dicho atraso inaceptable en términos políticos: «*la UE se arriesga a perder las oportunidades que brinda la IA, lo que la abocaría a una fuga de cerebros y a convertirse en consumidora de soluciones desarrolladas en otros lugares*»[76].

En definitiva, el contexto sociopolítico en el que nos movemos parece declinar claramente la balanza de los beneficios sobre los riesgos[77]. No obstante, autores como Cabitza tienen dudas sobre el "balance neto" en el uso de estas tecnologías. Bajo su punto de vista, en estos momentos es difícil predecir si el balance neto del uso de las tecnologías de IA en sectores complejos será positivo o negativo, siendo igualmente complicado prever qué puede contribuir en una dirección o en otra, y la propia definición de "éxito" o "fracaso". Ofrece algunos ejemplos como el impacto en el empleo, donde un impacto negativo como el aumento del desempleo -argumento recurrente al hablar de automatización como veremos a continuación-, puede verse solapado por otro positivo, como la reconversión de la mano de obra en ámbitos más creativos o gratificantes. Es difícil, dice, entender si

[75] Cowls, «'AI for Social Good': Whose Good and Who's Good? Introduction to the Special Issue on Artificial Intelligence for Social Good».

[76] Vid. Comisión Europea, Comunicación de la Comisión «Plan coordinado sobre la inteligencia artificial». Bruselas, 7.12.2018. COM(2018) 795 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=COM:2018:795:FIN

[77] Basta comparar los considerandos 3 y 4 de la propuesta de Reglamento AIA para observar que los potenciales y múltiples beneficios que la Comisión observa en estas tecnologías merecen mayor atención que sus riesgos: *(3) La inteligencia artificial es un conjunto de tecnologías de rápida evolución que puede generar un amplio abanico de beneficios económicos y sociales en todos los sectores y actividades sociales. El uso de la inteligencia artificial puede proporcionar ventajas competitivas esenciales a las empresas y facilitar la obtención de resultados positivos desde el punto de vista social y medioambiental en los ámbitos de la asistencia sanitaria, la agricultura, la educación y la formación, la administración de infraestructuras, la energía, el transporte y la logística, los servicios públicos, la seguridad, la justicia, la eficiencia de los recursos y la energía, y la mitigación del cambio climático y la adaptación a él, entre otros, al mejorar la predicción, optimizar las operaciones y la asignación de los recursos, y personalizar las soluciones digitales que se encuentran a disposición de la población y las organizaciones; (4) Al mismo tiempo, dependiendo de las circunstancias de su aplicación y utilización concretas, la inteligencia artificial puede generar riesgos y menoscabar los intereses públicos y los derechos que protege el Derecho de la Unión, de manera tangible o intangible.*

la potencia computacional de estos sistemas puede ayudarnos a afrontar la inminente crisis climática -por ejemplo, encontrando soluciones más eficaces, produciendo mejores modelos predictivos o haciendo más eficientes muchos procesos industriales-; o, por el contrario, si estas "megamáquinas" empeorarán el balance neto de emisiones dadas sus necesidades energéticas -debido a las emisiones de gases de efecto invernadero producidas a lo largo del ciclo de vida de los sistemas de IA-[78].

No obstante, estas primeras reflexiones sobre la "imagen pública" de estas tecnologías y su "balance neto", no son más que un análisis de brocha gorda. Analizar en bruto el contexto político, social y económico en el que se desarrollan estas tecnologías sería prácticamente inabarcable. Por ello, y dado que este apartado tiene un objetivo introductorio en la investigación, se han seleccionado tres fenómenos o perspectivas desde las que la literatura ha venido analizando dicho contexto, a saber: la automatización, la sociedad de los datos y la regulación algorítmica.

2.1. Automatización

Con automatización se hace habitualmente referencia a la sustitución de seres humanos por máquinas en determinadas tareas o procesos de toma de decisiones. Mientras que la automatización hacía referencia en su sentido más clásico a la sustitución de la fuerza bruta humana, la automatización digital actual enfatiza especialmente sobre la sustitución del pensamiento humano[79]. Por lo general, aumentar la productividad a través de la automatización implica la necesidad de menos humanos para alcanzar el mismo resultado[80]. Con todas las consecuencias que ello conlleva.

La automatización y sus consecuencias sociales se han analizado desde la primera Revolución Industrial. Estos análisis se han centrado muy especialmente en el impacto de la automatización en las relaciones laborales[81], aunque el impacto social y económico de la introducción de nuevas formas de automatización es mucho más hondo y trasciende el

[78] Floridi y Cabitza, *L'intelligenza artificiale. L'uso delle nuove macchine*.

[79] Bostrom y Yudkowsky, «The ethics of artificial intelligence».

[80] Müller, «Ethics of Artificial Intelligence and Robotics».

[81] Los procesos de automatización impulsados en la era de digitalización actual son analizados como parte de la 4ª Revolución Industrial. Vid. Poquet Catala, «Cuarta revolución industrial, automatización y afectación sobre la continuidad de la relación laboral».

objeto de dichos análisis[82]. Del mismo modo, la resistencia popular a la automatización de los procesos de producción no ha sido nunca un simple rechazo a la máquina o respuesta tecnófoba[83], sino una contestación a las formas de organización del trabajo y de sus réditos que se imponían en el ámbito sociolaboral a partir de la introducción de las máquinas en el proceso productivo.

La capacidad de esta clase de modelos algorítmicos para automatizar tareas que no necesariamente tienen un carácter rutinario y repetitivo, sino tareas más complejas como la conducción, así como muchas actividades mentales que durante mucho tiempo se han considerado no susceptibles de ser automatizadas, como la contabilidad, la planificación, la investigación jurídica o el diagnóstico médico, ha llevado a considerar que estas tecnologías podrían tener un impacto más profundo y a más largo plazo en el futuro del trabajo humano que las revoluciones industriales anteriores[84]. No parece sencillo realizar una predicción de este calado, aunque sí resulta llamativo que el presente proceso de automatización tenga su efecto en profesiones liberales, habitualmente excluidas del análisis y efectos de la automatización.

Cualquier proceso de automatización, incluso de tareas aparentemente simples, debe ser evaluado teniendo en cuenta el impacto que produce sobre el conjunto del entorno que suele ser más profundo de lo que intuitivamente percibimos.

De Sio y van Wynsberghe analizaron la automatización de una tarea simple como la recogida de una muestra de orina para analizar la presencia (o no) de toxinas de quimioterapia en niños sometidos a este tratamiento. Entre las conclusiones de este estudio hallaron que, desde el punto de vista de la eficacia, podría decirse que la recogida de la muestra la realiza mejor un robot; además, la recogida de la muestra puede ser embarazosa para el paciente y peligrosa para la salud del personal de enfermería, y eliminar al ser humano de la tarea evitaría estos otros riesgos no deseados. No obstante, esta tarea también forma parte de una actividad asistencial más amplia, cuyo objetivo es

[82] Más amplio es el análisis del European Group on Ethics in Science and New Technologies (EGE). Vid. European Group on Ethics in Science and New Technologies, «Future of Work, Future of Society».

[83] En este sentido ha contribuido, por ejemplo, la caricaturización del ludismo como un movimiento tecnófobo fruto de la ignorancia y el miedo. Cavero Garcés, «La cólera de Ludd y Swing. El luddismo industrial y agrario en el primer tercio del siglo XIX», 41.

[84] Santoni de Sio, Almeida, y van den Hoven, «The future of work: freedom, justice and capital in the age of artificial intelligence», 3.

el mantenimiento o el establecimiento del bienestar del paciente; lo cual parece requerir, entre otras cosas, que el paciente pueda acceder a cierta interacción con una o varias personas cuidadoras y que el paciente pueda establecer una relación de confianza con ellas.

Por lo tanto, incluso un proceso relativamente sencillo como la recogida de muestras de orina puede no estar totalmente automatizado y se requiere la presencia humana a lo largo del proceso, aunque posiblemente no en el momento mismo de la recogida[85]. Lo cual invita a repensar la manida problemática del desempleo tecnológico como un efecto secundario de una política y una economía de la tecnología informadas por el reconocimiento de diferentes valores como constitutivos de diferentes actividades; y que promueven el diseño de tecnologías e instituciones que reflejan los valores apropiados de diferentes esferas de la vida[86].

En este trabajo trataré de indagar en los efectos de la automatización sobre la toma de decisiones, con especial atención a los procesos en los que parte de la actividad de un operador humano queda automatizada por el uso de tecnologías como un soporte o apoyo a la toma de decisiones que le corresponde. Con carácter general, puede decirse que la automatización puede ayudar a centralizar y aumentar el control sobre múltiples procesos para los responsables, al tiempo que limita el poder discrecional de los operadores humanos en la cadena de toma de decisiones[87]. Por todo ello, tiene una gran relevancia analizar desde la ética-jurídica el desplazamiento o no de la responsabilidad sobre un proceso entre el operador humano y el responsable que decide sobre dicha automatización y sobre el resto de *responsables* que forman parte del ciclo de vida del sistema automatizado.

2.1. La sociedad de los datos o la *datificación* de la sociedad

Resulta habitual oír hablar de los datos como el nuevo petróleo. Los datos son, de acuerdo con quienes recurren a este símil, una nueva fuente de riqueza que impulsará el desarrollo

[85] Vid. Santoni de Sio F and van Wynsberghe A, 'When Should We Use Care Robots? The Nature-of-Activities Approach' (2016) 22 Science and Engineering Ethics 1745

[86] Santoni de Sio, Almeida, y van den Hoven, «The future of work: freedom, justice and capital in the age of artificial intelligence», 11.

[87] Noorman, «Computing and Moral Responsibility». Aunque tampoco puede descartarse que ese impacto sea en ocasiones ambivalente respecto de la autonomía humana, European Group on Ethics in Science and New Technologies, «Future of Work, Future of Society», 68.

industrial 4.0 en la era de la información[88]. No obstante, el uso extensivo y ubicuo de los datos generados por la actividad humana tiene una lectura socioeconómica bastante más compleja.

La datificación se ha definido como la transformación de la acción social en información cuantificada en línea que permite el seguimiento en tiempo real y el análisis predictivo[89]. Comprende, además, dos procesos: la transformación de la vida humana en datos a través de procesos de cuantificación, y la generación de diferentes tipos de valor a partir de los datos[90]. Y conlleva, a su vez, la legitimidad para acceder, comprender y controlar el comportamiento de las personas sobre, primero, la creencia generalizada en la objetividad del seguimiento de todo tipo de comportamiento humano y social a través de las tecnologías de la información; y segundo, la confianza en los agentes que recogen, interpretan y comparten los (meta)datos obtenidos de las redes, plataformas y tecnologías extractivas[91]. En otras palabras, el *big data* se presenta a sí mismo como científica y políticamente neutro, impregnado de un *aura de verdad, objetividad y precisión*[92].

La sociedad de los datos funciona, en gran medida, sobre bases matemáticas. No obstante, aunque se han desarrollado códigos de conducta y directrices para las autoridades estadísticas, las organizaciones que dependen de los procesos de Big Data no están obligadas a cumplir estos códigos y principios, lo cual se traduce en que muchos errores y sesgos persisten tanto en los conjuntos de datos, como en los algoritmos y sus resultados[93]. La más que explorada relación entre las grandes corporaciones tecnológicas y la vulneración de los más elementales estándares éticos ha revelado una de las

[88] Desde luego estos análisis suelen obviar también la contrapartida de este símil, y es que la huella ecológica de una sociedad datificada excede los límites biofísicos del planeta. Vid. Nardi et al., «Computing within limits».

[89] Mayer-Schönberger y Cukier, *Big Data: A Revolution that We Transform How We Live, and Think*, 30.

[90] Mejias y Couldry, «Datafication», 2.

[91] van Dijck, «Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology», 198.

[92] Boyd y Crawford, «Critical Questions for Big Data», 663.

[93] van der Sloot y van Schendel, «Procedural law for the data-driven society», 306. Es cierto que la ética ha ocupado un lugar central en el debate público sobre la ciencia de datos, no obstante, Green advierte de que centrándose en la ética sin unos principios normativos o deliberativos sólidos, ignorando o dando por sentado que las principales cuestiones políticas están resueltas, la ciencia de datos solidifica las estructuras sociopolíticas existentes y estrecha nuestra perspectiva sobre la posibilidad y deseabilidad de un cambio social más profundo. Vid. Green, «Data Science as Political Action: Grounding Data Science in a Politics of Justice».

principales problemáticas en la gobernanza de la datificación de la sociedad. Además, no se trata únicamente de rigor científico, de aplicar unos mayores estándares, sino que la propia legitimidad científica que se otorga a la denominada ciencia de datos contiene también fundamentos cuestionables.

Es decir, la revolución de los datos ha encontrado en la llamada "ciencia de datos" una fuente de autoridad para su expansión. En su artículo "Data Science as Machinic Neoplatonism", McQuillan explica cómo la ciencia de datos no es un simple método, sino una idea organizativa del mundo basada en que las cualidades susceptibles de ser datificadas revelan -a través de la computación y la correlación- un orden matemático del mundo, una capa de realidad, no aprehensible a través de la experiencia[94]. Es decir, los acontecimientos en la ciencia de los datos se constituyen no a partir de la experiencia, sino de los rastros de dicha experiencia que pueden ser expresados en datos -*inputs*-, cuya consecuencia es el desplazamiento del significado lejos de la aprehensión directa[95].

Como consecuencia de todo ello, la ciencia de datos pretende mantenerse al margen del mundo que observa o manipula y sitúa a los datos en un lugar ontológicamente superior, como si existiesen más allá del contexto en el que se extraen[96]. Desde luego no actuamos como meros receptores de los análisis algorítmicos, sino que somos los datos que se utilizan para tomar decisiones sobre nosotros. Incluso cuando éstos se anonimizan o codifican, detrás de dichos análisis algorítmicos no hay otra cosa que datos generados a partir de la actividad humana.

Reducir la actividad humana a lo datificable, tratar de datificar toda la actividad posible u otorgar una importancia elemental a lo datificado sobre lo no datificado, son cuestiones que revelan problemáticas más allá de quién posee los datos y cómo los utiliza o de si los datos son de calidad o están sesgados[97]. El riesgo de que la ciencia de datos nos alumbre

[94] Por dicha razón McQuillan habla de la ciencia de datos como una nueva forma de neoplatonismo.

[95] McQuillan, «Data Science as Machinic Neoplatonism».

[96] Hoffmann, «Making Data Valuable: Political, Economic, and Conceptual Bases of Big Data», 211.

[97] Méndez relata lo siguiente: (…) vieron cómo la investigación social merma y modifica las experiencias vividas cuando estas se intentan medir de tal modo que puedan ser procesadas por algoritmos. Por ejemplo, las evaluaciones tienden a tener en cuenta factores de riesgo tales como la asistencia a la escuela, o las denuncias de maltrato y abuso en el hogar, fácilmente cuantificables. Pero no saben considerar factores positivos como las redes de apoyo familiar extendidas, el compromiso social o la implicación en el barrio, pues estos factores dependen de un conocimiento contextual y de una información no estructurada (…) al final se acaba midiendo lo fácil y relegando lo difícil. Mendez, *Ciencia sin ficción. Cinco historias*, 228.

con una falsa claridad, simplificadora y reduccionista de la realidad, es evidente[98]. Y, por supuesto, puede y debe levantarse esa aura de *verdad, objetividad y precisión*, sin descartar el valor de la ciencia de datos y su capacidad para generar proposiciones válidas sobre nuestro mundo.

El impacto de una sociedad datificada obliga al ordenamiento jurídico a determinar las formas y estándares bajo los que dichos procesos de datificación son aceptables, valiosos y realizables y, muy particularmente, a determinar aquellos ámbitos de la vida social que deben quedar al margen de la datificación.

2.2. Regulación algorítmica

La regulación algorítmica es la más reciente de las perspectivas desde las que se han analizado el impacto social, económico y político de los modelos algorítmicos aquí descritos, y ha sido definida por Yeung en los siguientes términos: «*decision-making systems that regulate a domain of activity in order to manage risk or alter behaviour through continual computational generation of knowledge from data emitted and directly collected (in real time on a continuous basis) from numerous dynamic components pertaining to the regulated environment in order to identify and, if necessary, automatically refine (or prompt refinement of) the system's operations to attain a pre-specified goal*»[99]. De forma más amplia, podemos hablar de la regulación algorítmica como la conjunción entre la automatización y la datificación de la sociedad que genera

[98] Allo explica cómo se produce la tendencia a utilizar conceptos imprecisos como si fueran nítidos en la ciencia de datos: *we replace a question of interest ('is this a cat?') that may not have a determinate answer with a proxy-problem that does have a determinate answer ('is this pattern present?') and can therefore be algorithmically resolved, it is tempting to confuse our ability to correctly solve the proxy-problem with our ability to provide a correct answer to the actual problem*. Allo, «Mathematical Values And The Epistemology Of Data Practices», 20-23.

[99] Yeung, «Algorithmic regulation: A critical interrogation». Un ejemplo de regulación algorítmica siguiendo esta definición es la regulación que plataformas como Uber utilizan en su actividad, entre otros, para regular la actividad de los conocidos como *drivers*. Eyert et al. desarrollaron un estudio de caso sobre esta clase de regulación algorítmica. Mostraron cómo la aplicación reúne los aspectos del sistema que se consideran relevantes para regularlo (dimensión de representación), también mostraron las elecciones del regulador -Uber- de los estados deseados y las formas de hacerlos cumplir (dirección) y los intentos de desplazar a los regulados -drivers- hacia un estado deseable (intervención). El análisis, aunque muestra algunos rasgos distintivos de la regulación algorítmica, permite a los autores confrontar la narrativa de una "economía colaborativa" no jerárquica con las formas reales en que Uber regula la actividad de sus drivers, que con frecuencia se asemejan a los procedimientos de las empresas más convencionales. Vid. Eyert, Irgmaier, y Ulbricht, «Extending the framework of algorithmic regulation. The Uber case», 13-15.

ecosistemas algorítmicos en los que la realidad se percibe, se interpreta y se genera a partir de procesos computacionales[100].

El uso de algoritmos predictivos para regular determinadas actividades se enfrenta a las limitaciones que el término "aprendizaje" tiene cuando hacemos referencia al aprendizaje automático. Los datos son una foto del pasado y las correlaciones establecidas a partir de los mismos una forma de optimizar y ver dicho pasado, nunca el futuro que pretenden predecir. En palabras de Hildebrandt, mientras que podemos desarrollar muchos futuros-presentes (predicciones, imaginaciones, anticipaciones), solo tenemos un futuro-presente[101]. McQuillan explica el potencial de las características computacionales de la ciencia de datos para producir distorsiones y prejuicios a gran escala, como resultado de la fusión entre matemáticas y cultura: «*The algorithmic eye is not ocular but oracular. (…) The modulation of the present in the name of an algorithmic vision of the future forces us to ask what elements from the past are being projected in to that future, and hence in to the now*»[102].

Las predicciones no describen la realidad. Señala Véliz que una predicción es una conjetura, y en ella se incorporan todo tipo de apreciaciones y sesgos subjetivos sobre el riesgo y los valores. Puede haber previsiones más o menos exactas, sin duda, pero la relación entre probabilidad y realidad es mucho más tenue y éticamente problemática de lo que algunos suponen[103]. Añade que son tres los principales problemas éticos que surgen en el uso de algoritmos predictivos:

El primero, es que realizar previsiones sobre el comportamiento humano al igual que hacemos previsiones sobre el tiempo, reduce a las personas a cosas[104]. Es necesario

[100] Vid. Katzenbach y Ulbricht, «Algorithmic governance».

[101] Y añade que, teniendo en cuenta el impacto de las mismas, es posible que queramos ser prudentes a la hora de predecir. Hildebrandt, «Code-driven Law: Freezing the Future and Scaling the Past», 75.

[102] A su modo de ver, es muy probable que la regulación algorítmica provoque resultados distorsionados a escala, incluso de forma no intencionada y a pesar de los esfuerzos por mitigar los sesgos de los datos. En este mismo texto aboga por sustituir los estrechos márgenes de un aprendizaje automático que induce a la indefensión y la ausencia de reflexión a través de la manipulación de datos a escala, por una idea más amplia de aprendizaje que aumente la confianza de las personas en sí mismas y en su capacidad para resolver problemas. Por ende, haciendo uso del aprendizaje automático en tanto impulse esta idea más amplia de aprendizaje. Vid. McQuillan, «Algorithmic paranoia and the convivial alternative», 3-10.

[103] Veliz, «If AI Is Predicting Your Future, Are You Still Free? Part of being human is being able to defy the odds. Algorithmic prophecies undermine that». Disponible en: https://www.wired.com/story/algorithmic-prophecies-undermine-free-will/

[104] Veliz.

remarcar que los algoritmos solo pueden aprender aspectos del contexto que pueden ser matematizados -datificados-, por ende, los intentos algorítmicos por explicar lo que sucede reduce el sistema a ciertos elementos constitutivos y sus interacciones. Si, además, aplicamos este aprendizaje a cuestiones sociales, los atributos datificados serán maquinizados como individuales e innatos, eliminando las causas sociales comunes a éstos[105]. Todo ello nos remite a las problemáticas ya abordadas al hablar de la datificación de la sociedad.

El segundo es que, tratando a las personas como cosas, estamos creando profecías autocumplidas. Y es que los análisis predictivos, en realidad, están "creando" en parte la realidad que pretenden predecir -lo cual crea a su vez un bucle difícilmente rebatible desde el régimen jurídico actual[106]-. La ciencia de datos no es exclusivamente una nueva forma de conocimiento, sino que actúa directamente sobre la producción del conocimiento como consecuencia de la computación[107]. Las predicciones afectan a la anticipación de las interacciones y provocan un (re)ajuste de las acciones que contribuye a un presente-futuro diferente que si no se hubieran considerado dichas predicciones[108]. En particular, el riesgo de generar profecías autocumplidas a partir de la eliminación de las causas sociales que determinan los atributos sobre los que se generan las predicciones, parece atentar contra los principios fundacionales del Estado Social y Democrático de Derecho.

Por último, añade Véliz, el uso extensivo de la analítica predictiva nos roba la oportunidad de tener un futuro abierto en el que podamos marcar la diferencia, y esto puede tener un

[105] McQuillan, «The Political Affinities of AI», 165.

[106] Van der Sloot y van Schendel exponen este bucle con los siguientes ejemplos: *(6) It is difficult under the current legal regime to complain about positive things that did not happen due to data analytics. Suppose the police decide, on the basis of predictive policing, to patrol mainly in the southern district of a city and less so in the northern district. One question is whether an inhabitant of the southern district can file a complaint because she believes that the database, used for these predictions, is biased; another question is whether an inhabitant of the northern district can do so because she wants more surveillance in her neighbourhood. Most jurisdictions do not provide for such a possibility.; (7) An additional problem could emerge when the police decide to pay special attention to, for example, drug criminals and search available databases for clues, restricting the search queries to inhabitants of a particular neighbourhood, with a high density of people with a migration background. Suppose the results obtained, combined with additional evidence, subsequently lead to an arrest and criminal procedure. Can a person against whom irrefutable evidence has subsequently been found that she is dealing in drugs object to the evidence produced because the initial search was biased?*. van der Sloot y van Schendel, «Procedural law for the data-driven society», 306.

[107] McQuillan, «Data Science as Machinic Neoplatonism», 262.

[108] Hildebrandt, «Code-driven Law: Freezing the Future and Scaling the Past», 75.

impacto destructivo en la sociedad en general[109]. Estos problemas nos enfrentan a la "irresoluble tensión" entre la práctica de la predicción del comportamiento humano y la creencia en el libre albedrío como parte de nuestra vida cotidiana[110]. El reduccionismo de esta clase de tecnologías y su afán de anticipación empujan, además, a la tentadora eliminación de lo indeseable[111]. Cabitza habla de "esclerosis epistémica", esto es, el riesgo de perder el hábito de explorar lo desconocido y gestionar, también en términos de conciencia, tolerancia e incluso apreciación, la incertidumbre que afecta a todas nuestras evaluaciones, estimaciones y predicciones[112].

Tampoco faltan voces en sentido contrario, esto es, voces que conciben la regulación algorítmica como una oportunidad para el libre desarrollo de la personalidad. Para Domingos, si aprendemos cómo manejar el aprendizaje automático desde una perspectiva individual, la regulación algorítmica no determinará nuestro futuro más que otras tecnologías. Nuestra tarea es, a su parecer, entender estos algoritmos y moldearlos para que su aprendizaje se adapte a nuestras necesidades[113].

En cualquier caso, y para finalizar, conviene advertir que la regulación algorítmica no es un poder algorítmico en sí. Aunque en los términos expresados la regulación algorítmica tenga unos efectos determinados y particulares sobre la sociedad, los algoritmos no ejercen un poder sobre ésta, lo hacen las personas. Cuando se añade el apelativo "algorítmico" al análisis del contexto, éste no puede enmascarar las estructuras de poder -humano- que generan las condiciones para que se tomen las decisiones -ya sean prescritas por humanos o por ordenadores en última instancia-[114].

---

[109] Veliz, «If AI Is Predicting Your Future, Are You Still Free? Part of being human is being able to defy the odds. Algorithmic prophecies undermine that».

[110] Veliz.

[111] McQuillan, «The Political Affinities of AI», 165.

[112] Floridi y Cabitza, *L'intelligenza artificiale. L'uso delle nuove macchine*, 133-34.

[113] Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, 263-64. Para mí, esta visión edulcorada sobre nuestra capacidad individual para "moldear" la regulación algorítmica es difícil de sostener. Hemos de tener en cuenta que el entrenamiento de los algoritmos se produce sobre datos de los que no tenemos control, su aprendizaje responde a variables que los usuarios finales no pueden discutir (incluso en caso de que los conozcan y entiendan) y, por tanto, el control sobre los resultados algorítmicos que afectan a una persona puede ser poco más que una ilusión para el conjunto de la sociedad.

[114] Katz, «Manufacturing an Artificial Intelligence Revolution», 15.

3. El papel del Derecho en este contexto

En este apartado se expondrán unas líneas muy generales acerca de cuál ha sido el papel del ordenamiento jurídico ante la proliferación de estos modelos algorítmicos para la toma de decisiones y elaboración de perfiles en tantos ámbitos de nuestra vida, desde lo cotidiano al ámbito público de toma de decisiones.

En demasiadas ocasiones la respuesta del legislador ante el desarrollo exponencial de nuevas tecnologías ha resultado tardío, condicionado e insuficiente: «*Tardío porque los daños a bienes jurídicos importantes ya se han producido y seguirán produciéndose en cuanto las propias tecnologías albergan en el seno de su configuración la capacidad de atentar contra la privacidad. Condicionado porque reacciona sólo antes casos puntuales. Insuficiente porque las soluciones encontradas no parten de un análisis real de la raíz del problema, limitándose a corregir déficits concretos. Y siempre focalizado a responder a las consecuencias no a las causas de los problemas que se van poniendo de relieve*»[115].

En el caso particular de los sistemas tecnológicos referidos en esta investigación, Boix expone en su análisis que la reacción inicial de nuestros ordenamientos jurídicos ha sido tratar de obviar la existencia de sistemas que supongan un cambio cualitativo respecto del estado tecnológico precedente y, por tanto, merezcan una reconsideración de la respuesta jurídica aplicable en distintos ámbitos del ordenamiento -ello cuando no se ha optado directamente por la vía de la prohibición de su uso-[116].

Si, tal y como recogía el Consejo de Europa, la introducción de complejos algoritmos y programas informáticos para transformar los datos masivos en un recurso para la toma de decisiones plantea muchos problemas de protección de datos[117], lo cierto es que el Reglamento General de Protección de Datos, a pesar de que su entrada en vigor data de 2016 y su plena aplicabilidad de 2018, no ha conseguido abordar de forma satisfactoria los retos que estos modelos algorítmicos plantean. En los trabajos preparatorios, el contenido del artículo 22, sobre la toma de decisiones automatizadas, se consideraba

[115] de la Mata y Barinas Ubiñas, «La privacidad en el diseño y el diseño de la privacidad, también desde el Derecho Penal», 261.

[116] Boix, «Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones», 230.

[117] Consejo de Europa. Consultative Committee of Convention 108 y Council of Europe, «Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data».

crucial y la elaboración de perfiles encontró también un lugar destacado en el desarrollo de la norma[118]. No obstante, y sobre ello me extenderé en el capítulo correspondiente[119], hasta el momento la regulación de la toma de decisiones automatizada y la elaboración de perfiles en el RGPD no está cumpliendo un papel central en la aplicación de la normativa[120].

Además, algunas de las problemáticas acaecidas cuestionan los mismos fundamentos de la protección de datos personales, como el hecho de que el uso de datos no personales, especialmente en diseño y desarrollo está muy extendido o la aparición de cuestiones sociales que trascienden el interés individual[121]. Respecto de esta última cuestión, podemos encontrar un consenso en la idea de que el enfoque clásico de la privacidad y la protección de datos se ve superado por la dimensión colectiva del uso masivo de datos y su aprovechamiento algorítmico[122], el daño que causan no tiene una mera naturaleza privada y su dimensión pública (explicación) trasciende del concepto clásico de grupo "lo que pasa en Las Vegas se queda en Las Vegas"[123].

Aquí existe una tensión manifiesta entre quienes proponen una revisión de los derechos a la privacidad y a la protección de datos en su dimensión colectiva[124], frente a quienes

[118] Korff, «Comments on Selected Topics in the Draft EU Data Protection Regulation», 19.

[119] Vid. Capítulo 2. Toma de decisiones automatizada en el RGPD: el artículo 22 en la unidad de cuidados intensivos. Diagnóstico y propuestas terapéuticas para su recuperación.

[120] Sí ha reunido, no obstante, un interés destacado en la literatura jurídica.

[121] van der Sloot y van Schendel, «Procedural law for the data-driven society», 306-7. En este sentido, el RGPD excluye de su aplicación el tratamiento de datos no personales, y en la propia definición de dato personal (art. 4(1)) y en su modelo de protección individual de los derechos del interesado (arts. 12 a 23) parece hacer prevalecer el interés individual como fundamento principal de la protección de datos -aunque este punto me parece rebatible-.

[122] Quizás esta crítica sea extensible al conjunto de la regulación vigente de aplicabilidad en este contexto: Current ethical debates about the consequences of automation generally focus on the rights of individuals. However, algorithmic processes – the major component of automated systems – exhibit a collective dimension first and foremost. This can only be addressed partially at the level of individual rights. For this reason, existing ethical and legal criteria are not suitable (or, at least, are inadequate) when considering algorithms generally. Jaume-Palasi y Spielkamp, «Ethics and algorithmic processes for decision making and decision support», 4.

[123] Loi y Christen, «Two Concepts of Group Privacy».

[124] No son pocas las voces de la doctrina que han señalado que el foco en la privacidad, entendida como el derecho individual al control sobre la información que concierne a uno mismo, falla a la hora de reconocer hasta qué punto muchos derechos civiles y políticos están enraizados en última instancia en una estructura social y política en la que la privacidad se debe entender también como un bien colectivo. Yeung, «Algorithmic regulation: A critical interrogation», 517.

abogan por abandonar este marco de referencia de la protección de datos como punto de partida normativa[125].

No obstante, esta investigación ahonda más en otro aspecto que, motivado por el paradigma de las tecnologías de IA, en palabras de Kuner et al. obliga a la revisión de los principios tradicionales de protección de datos y a la reflexión sobre los nuevos mecanismos de protección de datos: la consideración del papel de las personas en la supervisión de la tecnología[126]. Parece evidente que a medida que aumentan la velocidad, la precisión y el impacto de la IA, es probable que el papel de la supervisión humana también deba cambiar[127]. La consideración de la supervisión humana como un requisito de obligado cumplimiento para las tecnologías de IA de alto riesgo -en las propuestas regulatorias europeas mencionadas en el siguiente apartado- parecen confirmar la importancia de revisar jurídicamente este aspecto que, como veremos, está presente -aunque haya pasado prácticamente desapercibido- en el artículo 22 RGPD en forma de intervención humana.

Más allá del RGPD, también porque su alcance es limitado, uno de los principales retos del Derecho consiste en abordar una respuesta jurídica del ordenamiento adecuada para el diseño y desarrollo de estos sistemas. La normativa aplicable a estas fases iniciales y determinantes del ciclo de vida de los sistemas de IA está habitualmente pensada para el desarrollo de programas informáticos, productos o sistemas con características muy diferentes[128].

---

[125] Una propuesta muy interesante es "data pollution" de Ben-Shahar. Sostiene que, siguiendo el símil de que los datos son el petróleo del presente siglo, entonces la contaminación por datos es para nuestro siglo lo que la lo que la contaminación industrial fue para el anterior. A partir del análisis del daño eminentemente social que producen estas prácticas y del fracaso de los marcos normativos existentes (muy particularmente el de la privacidad y protección de datos), desarrolla el concepto de "data pollution" y propone un marco de regulación inspirado en el control de la contaminación para hacer frente a este fenómeno. Vid. Ben-Shahar, «Data Pollution».

[126] Kuner et al., «Expanding the artificial intelligence-data protection debate», 291.

[127] Kuner et al., 291.

[128] Por ejemplo, y entre otras, en materia de aviación el Reglamento (CE) n.º 300/2008 del Parlamento Europeo y del Consejo, de 11 de marzo de 2008, sobre normas comunes para la seguridad de la aviación civil; en materia de vehículos los Reglamentos n.º 168/2013 del Parlamento Europeo y del Consejo, de 15 de enero de 2013, relativo a la homologación de los vehículos de dos o tres ruedas y los cuatriciclos, o n.º 167/2013 del Parlamento Europeo y del Consejo, de 5 de febrero de 2013, relativo a la homologación de los vehículos agrícolas o forestales; o en materia sanitaria los Reglamentos 2017/745 del Parlamento Europeo y del Consejo, de 5 de abril de 2017, sobre los productos sanitarios o 2017/746 del Parlamento Europeo y del Consejo, de 5 de abril de 2017, sobre los productos sanitarios para diagnóstico in vitro.

La diferencia fundamental la encontramos en la uniformidad esperable en distintos contextos de dichos programas, productos y sistemas, mientras que el funcionamiento de los sistemas algorítmicos abordados en esta investigación, por la forma en que se produce su aprendizaje y exacerbado en ocasiones por determinadas características como la opacidad, contiene un grado de indeterminación mucho más alto en su implementación. Ello provoca que los sistemas de validación normativa establecidos para la comercialización o puesta en servicio de un determinado programa, producto o sistema, puedan ser insuficientes para garantizar la seguridad y el correcto funcionamiento de estas tecnologías.

Teniendo en cuenta que, hasta el momento, no disponemos de un marco jurídico común de aplicación a las fases de diseño y desarrollo de los modelos algorítmicos para la toma de decisiones automatizada -análogo a la regulación que la normativa de protección de datos establece para la fase de implementación-, el alcance de esta investigación es limitado en este aspecto.

Como ya apuntaba Cotino, a pesar de que podamos establecer determinados principios comunes regulatorios -para lo cual acudimos fundamentalmente a la normativa de protección de datos en este momento-, la tarea jurídica en este ámbito requiere determinar con precisión el marco jurídico-sectorial aplicable[129], así lo expresaba en relación a la regulación del tratamiento masivo de datos, también aplicable al uso de sistemas basados en dicho tratamiento: «*Para ello, una de las premisas jurídicas es determinar y en su caso diferenciar el tratamiento jurídico de la actividad de big data cuando se realiza ya por poderes públicos, ya por el sector privado. El marco jurídico puede ser diferente a partir de responsabilidad del estado, principio de legalidad, interés público, frente a la libertad de empresa y derechos en juego por el sector empresarial. Ya se trate del sector público o privado que realice acciones de big data, hay que plantearse la discrecionalidad o potestad para usar y tratar los datos masivos, la protección jurídica que tienen respecto de los métodos, tecnologías y resultados del big data, en especial, debe tenerse en cuenta la propiedad industrial así como la concurrencia de posibles*

[129] En esta línea, se ha realizado en paralelo a esta investigación un estudio sectorial sobre la aplicación de sistemas de IA en la asistencia sanitaria. Dicho estudio parte de las premisas generales y comunes desarrolladas aquí, no obstante, su publicación no se encuentra disponible por el momento.

obligaciones de transparencia y puesta a disposición de los datos abiertos para su reutilización»[130].

La aprobación en 2016 del Reglamento General de Protección de Datos y su regulación de la toma de decisiones automatizada y la elaboración de perfiles ha permitido abordar de forma transversal algunas de las problemáticas manifestadas por este fenómeno tecnológico y social[131]. Sin embargo, la capacidad de esta norma para actuar sobre las determinantes fases de diseño y desarrollo de los sistemas algorítmicos es muy limitada -dejando de lado otras limitaciones que afectan a la implementación y uso de los sistemas que serán abordadas más adelante-. Asimismo, la legislación sectorial que regula actualmente el diseño y desarrollo de dichos sistemas está, por lo general, diseñada para responder a las características de tecnologías distintas, lo cual provoca lagunas inasumibles desde un punto de vista jurídico-político como veremos.

3.1. Propuestas para la regulación europea de la inteligencia artificial

Esta preocupación ha tratado de ser abordada por las instituciones en los últimos años. De ahí que sea posible que, próximamente, dispongamos de un marco jurídico común más amplio que abarque también el uso de IA que no necesariamente se base en el tratamiento de datos personales, o cuyo uso no esté centrado en realizar inferencias sobre personas particulares –a pesar de incluir tratamiento de datos personales–, como es el caso de la conducción autónoma o de los sistemas de armas autónomos letales (SAAL). Tanto para las fases de diseño y desarrollo, como para su posterior implementación. Un marco jurídico que, en cualquier caso, parece que incluirá los sistemas de toma de decisiones automatizada basada en la elaboración de perfiles personales que es objeto de estudio en esta investigación -siempre que dichos sistemas representen un nivel de riesgo mínimo-.

En abril de 2021, la Comisión presenta la propuesta de Reglamento por el que se establecen normas armonizadas en materia de inteligencia artificial, actualmente en tránsito por el procedimiento legislativo ordinario. Se trata de una propuesta que distingue

[130] Cotino, «Big data e inteligencia artificial . Una aproximación a su tratamiento jurídico desde los derechos fundamentales», 136.

[131] En esta investigación veremos que ha sido la aplicación del RGPD lo que ha permitido que distintos tribunales de los EEMM hayan podido actuar frente al uso abusivo y los efectos nocivos de estos sistemas.

entre distintos niveles de riesgo de los sistemas de IA[132] y que centra el grueso de su regulación en los sistemas de alto riesgo a través del establecimiento de unos requisitos de obligado cumplimiento para el diseño y desarrollo de estos sistemas de forma previa a su implementación en el mercado europeo. Ahora bien, el contenido de esta propuesta no puede entenderse sin observar los antecedentes de la misma trabajados en el seno de las instituciones de la UE.

En 2018, la Comisión hizo pública la Estrategia europea sobre la IA[133] con el triple objetivo de potenciar la capacidad tecnológica e industrial de la Unión, prepararse para las transformaciones socioeconómicas que origina la IA y garantizar el establecimiento de un marco ético y jurídico apropiado para la misma. En paralelo, y como resultado del compromiso conjunto de los EEMM, la Comisión aprobaba también un Plan coordinado sobre la inteligencia artificial para la generación de una IA "made in Europe"[134], que fue renovado en 2021[135]. El grupo de expertos de alto nivel sobre inteligencia artificial (AI HLEG) fue designado y constituido para asesorar sobre la Estrategia europea en materia de IA, y sus resultados sirvieron de recursos para nuevas iniciativas políticas[136]. Entre otras, el 19 de febrero de 2020 se publica el Libro Blanco sobre la IA por la Comisión[137], coincidiendo a su vez con la publicación de la Estrategia europea de datos[138]. El Libro Blanco definió un ecosistema de confianza en el que debe promoverse un marco normativo para la AI que aborde las oportunidades y los riesgos de estas tecnologías. Más

[132] Establece así cuatro clases de sistemas: (1) prácticas prohibidas o de riesgo inaceptable; (2) sistemas de alto riesgo; (3) obligaciones de transparencia para determinados sistemas; y (4) resto de sistemas de IA.

[133] Comisión Europea, Comunicación de la Comisión « Inteligencia artificial para Europa». Bruselas, 25.4.2018. COM(2018) 237 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=COM%3A2018%3A237%3AFIN

[134] Comisión Europea, Comunicación de la Comisión «Plan coordinado sobre la inteligencia artificial». Bruselas, 7.12.2018. COM(2018) 795 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=COM:2018:795:FIN

[135] Comisión Europea, Comunicación de la Comisión «Fomentar un planteamiento europeo en materia de inteligencia artificial». Bruselas, 21.4.2021. COM(2021) 205 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=COM:2021:205:FIN

[136] Comisión Europea, Comunicación de la Comisión «Generar confianza en la inteligencia artificial centrada en el ser humano». Bruselas, 8.4.2019. COM(2019) 168 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=CELEX:52019DC0168

[137] Comisión Europea, Comunicación de la Comisión «Libro Blanco sobre la inteligencia artificial – un enfoque europeo orientado a la excelencia y la confianza». Bruselas, 19.02.2020. COM (2020) 65 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52020DC0065

[138] Comisión Europea, Comunicación de la Comisión «Una Estrategia Europea de Datos». Bruselas, 19.02.2020. COM (2020) 66 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A52020DC0066

recientemente, el Parlamento Europeo aprobaba una ambiciosa Resolución de 20 de octubre de 2020[139], en la que se incluía un anexo con una propuesta legislativa para la tramitación de un Reglamento sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas. Estas propuestas son abordadas en mayor profundidad a lo largo de la investigación.

En cuanto a las instituciones españolas, destacan la Carta de Derechos Digitales impulsada por la Secretaría de Estado de Digitalización e Inteligencia Artificial del Ministerio de Asuntos Económicos y Transformación Digital[140] y la Estrategia Nacional de Inteligencia Artificial (ENIA) elaborada por el Grupo de Trabajo Interministerial en Inteligencia Artificial, coordinado por el Ministerio de Ciencia e Innovación[141].

En definitiva, a pesar de los loables intentos por parte de las instituciones europeas y también de las instituciones españolas, parece razonable reconocer que el ordenamiento reacciona una vez más a trompicones frente a los efectos de unas tecnologías que ya son plenamente percibidos/sufridos por el conjunto de la sociedad[142]. No hay más que constatar que, entre los sistemas que ya están siendo utilizados de forma generalizada en nuestro entorno, podemos encontrar muchos de los sistemas considerados de "alto riesgo" -y que las instituciones consideran que necesitan de una regulación mucho más estricta- o incluso, en el más grave de los casos, sistemas que podrían ser prohibidos por estas propuestas.

4. Preguntas que se plantean en esta investigación

[139] 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)). Disponible en: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_ES.html

[140] Gobierno de España, *Carta de Derechos Digitales*, Plan de recuperación, transformación y resiliencia, 2021. Texto completo disponible aquí: https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf;

[141] Gobierno de España, Estrategia Nacional de Inteligencia Artificial (ENIA), Versión 1.0. Disponible en: https://portal.mineco.gob.es/RecursosArticulo/mineco/ministerio/ficheros/201202_ENIA_V1_0.pdf

[142] En este sentido, resulta pertinente explorar fórmulas distintas para el diseño de políticas públicas y, en particular, que permitan mejorar la capacidad de anticipación de forma estratégica, sensibilizar a la opinión pública sobre los problemas de la privacidad y generar un enfoque participativo en el diseño de las políticas venideras. Como ejemplo, vid. Rossi et al., «What if data protection embraced foresight and speculative design?»

Una vez realizada esta contextualización, de forma introductoria a los capítulos que forman parte de esta investigación, y sin ánimo de exhaustividad, se plantean en este apartado algunas de las preguntas que irán resolviéndose con el desarrollo de esta tesis doctoral:

¿Qué fases integran la toma de decisiones automatizada basada en la elaboración de perfiles y cuál es su relevancia normativa?

¿Cuál es el papel del elemento humano en las distintas fases de la toma de decisiones automatizada? ¿Es la plena automatización un escenario plausible y merecedor de interés jurídico?

¿Qué formas pueden adoptar los mecanismos de gobernanza basados en la participación humana en la fase de implementación de la toma de decisiones? Desde un punto de vista jurídico, ¿cuál es la diferencia entre conceptos como la supervisión humana y la intervención humana, entre otros?

¿Cómo pueden definirse los sesgos en la toma de decisiones automatizada desde el ámbito jurídico? ¿Cuál es la relación entre sesgos y discriminación? ¿Es la discriminación el único aspecto jurídico relevante en la producción de estos sesgos?

La literatura ha definido habitualmente estos modelos como cajas negras, ¿cuál es la relevancia jurídica de esta clase de opacidad? ¿Existen otras clases de opacidad? ¿Cuál es la relación que guardan las distintas clases de opacidad con la transparencia como principio normativo?

¿Qué relación guardan la protección de la privacidad y la protección de datos con las inferencias personales y la toma de decisiones automatizada? ¿Por qué es el RGPD el cuerpo normativo de mayor interés para el estudio jurídico del objeto de investigación?

¿Es satisfactorio el modelo regulatorio del RGPD para la gobernanza de la toma de decisiones basada en la elaboración de perfiles? Teniendo en cuenta que el artículo 22 RGPD se erige como pilar para dicha gobernanza y que, sin embargo, su aplicación es escasa y errática, ¿podemos hablar de un fracaso normativo?

Para la interpretación de la regulación de la toma de decisiones automatizada en el RGPD, la doctrina ha tomado como referencia el principio de transparencia y los derechos de

información y acceso, ¿es la transparencia el único fundamento normativo de este modelo de gobernanza, o es el fundamento más apropiado para la gobernanza de estos sistemas?

¿Cuáles son los distintos mecanismos de gobernanza basados en la intervención humana requeridos en la regulación de la toma de decisiones automatizada del RGPD? ¿Qué características y relevancia tiene cada uno de ellos en el ecosistema regulatorio del RGPD? ¿Cómo podemos entender la intervención humana significativa exigida en este contexto para los procesos de toma de decisiones automatizada?

¿Cuáles son las diferencias fundamentales de los remedios normativos previstos para la toma de decisiones basada únicamente en el tratamiento automatizada y la toma de decisiones no basada únicamente en el mismo? ¿Están justificadas estas diferencias y son apropiadas para asegurar la responsabilidad sobre el tratamiento automatizado, la posibilidad de observar el tratamiento de datos y sus efectos en virtud del principio de transparencia o la capacidad de la persona interesada para influir sobre el tratamiento o de contestar al mismo cuando no se ajusta al ordenamiento?

¿Para qué tipo de procesos de toma de decisiones automatizada basados en la elaboración de perfiles es obligatoria la realización de la evaluación de impacto de protección de datos?

¿En qué medida debe la evaluación de impacto de protección de datos analizar la inclusión y efectividad de la intervención humana en los procesos de toma de decisiones automatizada basados en la elaboración de perfiles?

¿Qué interés jurídico tiene entender la intervención humana conjuntamente como un mecanismo de gobernanza contenido en los derechos individuales reconocidos por el RGPD y como una medida organizativa exigida por el RGPD como desarrollo de la responsabilidad de cumplir y demostrar el cumplimiento normativo?

CAPÍTULO 1. MARCO TEÓRICO DE LA TOMA DE DECISIONES AUTOMATIZADA BASADA EN LA ELABORACIÓN DE PERFILES

CAPÍTULO 1. MARCO TEÓRICO DE LA TOMA DE DECISIONES AUTOMATIZADA BASADA EN LA ELABORACIÓN DE PERFILES

Para el desarrollo de esta investigación es indispensable partir de que los sistemas algorítmicos se despliegan en una amplia variedad de contextos sociales, algunos más problemáticos que otros. Ello, es evidente, supone que el análisis jurídico dista de un contexto a otro. No obstante, el análisis bibliográfico realizado arrojó pronto una serie de elementos comunes a distintos contextos, sobre los que pivota el acercamiento normativo a dichos sistemas. A su vez, en las distintas iniciativas para la regulación de los sistemas de IA de "alto riesgo" para una amplia variedad de contextos, estos elementos comunes son nuevamente destacados por la relevancia que adquieren a la hora de definir los modelos de gobernanza de estos sistemas. Todo ello, apunta a la utilidad de tratar de sistematizar estos elementos.

La sistematización que se realiza en el presente capítulo tiene por objetivo el desarrollo de un marco teórico que esclarezca el análisis jurídico de la toma de decisiones automatizada, sea cual sea el contexto de despliegue e implementación de los sistemas algorítmicos[143]. Este marco teórico era necesario para la realización de esta investigación, pero se convierte a su vez en un resultado por sí mismo, dado que puede servir de referencia a cualquier jurista que se acerque a la materia.

Para la realización de este marco teórico, en primer lugar, describo las distintas fases que comprende la toma de decisiones automatizada basada en la elaboración de perfiles, desde el diseño y desarrollo de los sistemas algorítmicos hasta su implementación en el mundo real. En esta descripción, con un carácter no exhaustivo, se hace referencia a las múltiples decisiones que los distintos agentes implicados pueden adoptar y a su relevancia normativa.

Posteriormente, a partir de los resultados del análisis bibliográfico realizado, se abordan tres aspectos particulares de este fenómeno, a saber: la participación o intervención humana en la implementación de los sistemas, los sesgos en la toma de decisiones y la opacidad de los sistemas algorítmicos. En los tres apartados se sigue una metodología

[143] A simple modo de recordatorio, volvemos sobre la definición de toma de decisiones automatizada basada en la elaboración de perfiles propuesta para la presente investigación: *proceso de toma de decisiones en el que, a partir del resultado de un sistema algorítmico que realiza una inferencia, clasificación o evaluación sobre una persona física, se adopta una decisión parcial o totalmente automatizada que afecta a la misma*.

similar; se realiza una tarea descriptiva de la intervención humana, los sesgos y la opacidad, distinguiendo entre distintas clases y perspectivas técnicas y normativas que se entrelazan, destacando especialmente estas últimas; y posteriormente, se describe cómo estos aspectos han sido abordados en las iniciativas de la UE para la regulación de los sistemas de inteligencia artificial, destacando aquí las cuestiones más discutidas en la literatura jurídica.

1. Fases de la toma de decisiones basada en la elaboración de perfiles

El primer objetivo en el diseño de este marco teórico de la toma de decisiones algorítmica es la de atomizar este proceso. Ahora bien, en esa división del todo se identificarán elementos comunes, básicos, de la toma de decisiones algorítmica; de tal modo que su nivel de abstracción sea suficiente para absorber avances en el estado del arte (técnico) que, evidentemente, seguirán produciéndose. Las fases que se han tomado como referencia han sido identificadas en las fuentes bibliográficas que se recogen a continuación.

Tomando el esquema utilizado por de Laat, la toma de decisiones basada en algoritmos de aprendizaje automático o *machine learning* a partir del tratamiento de datos masivos o *big data* se compondría de tres fases: (1) Recolección de datos – *data collection* – (2) Construcción del modelo – *model construction* – (3) Uso del modelo – *model use* –[144]. Dicho esquema estaría inspirado en las fases definidas por Zarsky como los tres segmentos de flujo de información y transparencia de los modelos algorítmicos predictivos, los cuales analizó a partir de la proliferación de prácticas gubernamentales predictivas basadas en el análisis de la información personal y potenciadas por la minería de datos[145]. Anteriormente, Schreurs et al. habían establecido ya un marco teórico para la elaboración y utilización de perfiles, que consistía también en tres pasos y cuya división coincide, en lo fundamental, con el marco anterior[146].

[144] de Laat, «Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?»

[145] (1) The collection of data and aggregation of datasets; (2) Data analysis; y (3) Usage stage. Zarsky, «Transparent predictions».

[146] (1) The First Step: Collection of Personal Data and Other Information to Construct Profiles; (2) The Second Step: The Construction of the Profile from Anonymous Data; y (3) The Third Step. The Application of the Group Profile. Schreurs et al., «Cogitas, Ergo Sum. The Role of Data Protection Law and Non-discrimination Law in Group Profiling in the Private Sector».

Más recientemente, en el ámbito del Derecho Público, destaca un estudio de de Fine y de Fine sobre utilización de Inteligencia Artificial en la toma de decisiones pública que sigue igualmente un esquema de tres fases, si bien, la primera fase se refiere al establecimiento de los objetivos del modelo, quedando la recolección de datos y la construcción del modelo integrados en una segunda fase de programación[147].

Coincidiendo con la necesidad de incorporar esa primera fase para el establecimiento de los objetivos del modelo, podemos tomar como referencia el sistema decisorio propuesto por Cobbe y Singh que incluye además una última fase de "investigación", dividiendo a su vez las fases en distintas etapas, y aboga por una comprensión de la toma de decisiones automatizada como un proceso socio-técnico, que implica tanto componentes humanos (organizativos) como técnicos, y que se producen antes de que se tome una decisión y se extienden más allá de la decisión propiamente dicha[148], una perspectiva totalmente compatible con la seguida en esta investigación.

Ahondando en la relevancia jurídica que puede adquirir esta atomización de la toma de decisiones algorítmica, el Parlamento Europeo, en la Resolución con recomendaciones destinadas a la Comisión sobre inteligencia artificial, robótica y tecnologías conexas[149], sigue en cierto modo este mismo proceso, considerando de relevancia tres fases que forman parte del título de la propuesta y define en su artículo cuarto: desarrollo, despliegue y uso[150]. Del mismo modo, la Comisión Europea en el Libro Blanco sobre la Inteligencia Artificial hace referencia a la necesidad de garantizar el cumplimiento de distintos requisitos normativos a lo largo de todo el 'ciclo de vida' de los productos y

[147] de Fine Licht y de Fine Licht, «Artificial intelligence, transparency, and public decision-making».

[148] Cobbe y Singh, «Reviewable Automated Decision-Making».

[149] 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)). Disponible en: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_ES.html

[150] Artículo 4: (…) f) «desarrollo», la construcción y el diseño de algoritmos, la escritura y el diseño de programas informáticos o la recopilación, el almacenamiento y la gestión de datos con el fin de crear o entrenar la inteligencia artificial, la robótica y las tecnologías conexas o de crear una nueva aplicación para la inteligencia artificial, la robótica y las tecnologías conexas existentes; (…) h) «despliegue», el funcionamiento y la gestión de la inteligencia artificial, la robótica y las tecnologías conexas, así como su comercialización o cualquier otra forma de puesta a disposición de los usuarios; (…) j) «uso»: toda acción relacionada con la inteligencia artificial, la robótica y las tecnologías conexas distinta del desarrollo o el despliegue;

sistemas de IA y concibe, al menos, una fase de diseño/desarrollo y una fase de uso[151]. En España, se alude también a dicho ciclo de vida de la IA en la guía publicada en febrero de 2020 sobre la adecuación al RGPD de los tratamientos que incorporan IA de la Agencia Española de Protección de Datos (AEPD, en adelante)[152]. La propuesta de Reglamento conocida como "Ley de Inteligencia Artificial" de la Comisión Europea (en adelante propuesta de Reglamento AIA o AIA), hace referencia a que los sistemas de alto riesgo deben funcionar de manera consistente durante *todo su ciclo de vida*, aunque a efectos regulatorios distingue exclusivamente entre la fase de desarrollo previa a la comercialización -para la que establece el grueso de las obligaciones jurídicas- y la fase de comercialización y utilización de los sistemas[153].

En la siguiente tabla se ilustran los modelos definidos por de Laat y Cobbe y Singh, junto con el modelo normativo por el que opta la Comisión en dicha propuesta:

<table>
<tr><th>DE LAAT</th><th colspan="2">COBBE Y SINGH</th><th>COMISIÓN EUROPEA</th></tr>
<tr><td>Fases</td><td>Etapas</td><td>Fases</td><td>Fases</td></tr>
<tr><td rowspan="2">(…)</td><td>Adquisición</td><td rowspan="2">Encargo</td><td rowspan="3">Diseño y desarrollo del modelo

-precomercialización-</td></tr>
<tr><td>Definición del problema</td></tr>
<tr><td>Recolección de Datos</td><td>Recolección de Datos</td><td>Construcción del modelo</td></tr>
</table>

[151] Comisión Europea, Comunicación de la Comisión «Libro Blanco sobre la inteligencia artificial – un enfoque europeo orientado a la excelencia y la confianza». Bruselas, 19.02.2020. COM (2020) 65 final.. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52020DC0065

[152] Como se verá, esta guía refleja las fases de concepción y análisis, desarrollo, explotación y retirada final, además de sus consiguientes subfases. Vid. Agencia Española de Protección de Datos (AEPD), «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción». Disponible en: https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf

[153] Siendo la «Introducción en el mercado» el momento que diferencia entre ambas fases, conforme al artículo 3(9) AIA. *«Introducción en el mercado»: la primera comercialización en el mercado de la Unión de un sistema de IA*.; a su vez, *«Comercialización»: todo suministro de un sistema de IA para su distribución o utilización en el mercado de la Unión en el transcurso de una actividad comercial, ya se produzca el suministro de manera remunerada o gratuita*. Comisión Europea, Propuesta de Reglamento por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial). Bruselas, 21.4.2021. COM(2021) 206 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex:52021PC0206

<table>
<tr><td rowspan="3">Construcción del modelo</td><td>Pre-procesado</td><td rowspan="3"></td><td rowspan="3"></td></tr>
<tr><td>Entrenamiento del modelo</td></tr>
<tr><td>Testeo del modelo</td></tr>
<tr><td rowspan="3">Uso del modelo</td><td>Despliegue</td><td rowspan="3">Toma de decisiones</td><td rowspan="3">Implementación del modelo

-postcomercialización-</td></tr>
<tr><td>Uso</td></tr>
<tr><td>Consecuencias</td></tr>
<tr><td rowspan="2">(…)</td><td>Auditoría</td><td rowspan="2">Investigación</td></tr>
<tr><td>Revelación</td></tr>
</table>

Tabla 1. Comparativa fases

A continuación, se procederá a definir brevemente las fases que se han tomado en consideración para la presente investigación, inspiradas fundamentalmente en los modelos descritos, y a su vez, entre otras cuestiones jurídicas relevantes, se destacará especialmente su conexión con la aplicación del Reglamento General de Protección de Datos (RGPD, en adelante), dado que el próximo capítulo aborda el análisis de esta normativa, así como con la propuesta de Reglamento AIA de la Comisión Europea, por su pretensión de regular de forma transversal el diseño y desarrollo de estos modelos. Ha de precisarse que estas fases no deben ser entendidas o utilizadas como compartimentos estancos o necesariamente sucesivos, su interrelación es constante y los límites que definen las mismas no son siempre nítidos.

1.1. Diseño y desarrollo del modelo

Tal y como arguye Djeffal, comprender el conjunto de decisiones contenidas en las fases de diseño y desarrollo es un paso esencial en lo que denomina la definición democrática de la tecnología: «*siempre que hay una alternativa, hay una elección*»[154]. A continuación,

[154] Añade: *From a democratic perspective, one must understand and highlight specific choices. These choices relate to architectures, applications and all other features of the technologies used. Whenever there is an alternative, there is a choice. Understanding choices also requires a democratic mindset that is open to several possibilities without automatically preferring certain outcomes. Computer scientists especially,*

abordaremos las múltiples alternativas que se dan tanto en el diseño -a la hora de definir el problema y de establecer los objetivos del modelo-, como en el desarrollo de un modelo -en la recolección de datos, su entrenamiento y validación-. Las decisiones adoptadas en estas fases tienen un carácter determinante respecto de la posterior fase de implementación[155], pudiendo incluso constreñir las posibilidades de control y cumplimiento de la normativa aplicable a esa fase posterior.

El establecimiento de unas normas armonizadas en estas fases es una prioridad para la Comisión Europea, para poder conseguir un mercado de IA con un nivel elevado de protección de los intereses públicos, como la salud y la seguridad, y de los derechos fundamentales[156]. En cualquier caso, ello no obsta para que la legislación vigente ya sea aplicable al diseño y desarrollo de modelos algorítmicos que pretendan comercializarse en el ámbito de la UE. La diferencia está en que la Comisión pretende ahora establecer un marco común para el diseño y desarrollo de estos sistemas complementario a la legislación sectorial aplicable[157], además de a sistemas destinados a ser utilizados en ámbitos particulares determinados ad hoc por la propia Comisión[158].

Del mismo modo, tal y como reconoce la exposición de motivos de la propuesta de Reglamento AIA, la aplicación del RGPD al diseño y desarrollo de estos modelos es determinante dada la necesidad de tratar datos para los fines del desarrollo -como veremos, tanto en la recolección, como en el entrenamiento y validación de los modelos-. No obstante, así como en la fase de implementación la aplicación del RGPD parece ineludible a la hora de realizar inferencias sobre personas determinadas, en las fases de diseño y desarrollo es habitual que el tratamiento de datos no se realice con datos personales y, por ende, se pueda eludir la aplicación del RGPD[159].

---

who are trained to achieve specific goals such as efficiency, regularly do not see behind the choices that maximize their preferred value. Djeffal, «AI, Democracy and the Law», 268.

[155] Por ejemplo, las decisiones sobre el modelo algorítmico escogido, más o menos complejo, puede limitar la interpretabilidad de la misma por el agente humano que tome las decisiones con apoyo algorítmico.

[156] Considerando 5 AIA.

[157] Entre otras mencionadas en el Anexo II AIA; legislación sobre juguetes, embarcaciones de recreo y a las motos acuáticas, ascensores o productos sanitarios.

[158] Recogidas en el anexo III conforme al artículo 6(2) AIA.

[159] Lo cual hace más pertinente si cabe, la necesidad de una regulación específica para las fases de diseño y desarrollo de estos modelos.

1.1.1. Diseño: Definición del problema y establecimiento de los objetivos del modelo

Esta fase, dentro del diseño y desarrollo de los modelos, engloba las decisiones de diseño que tienen que ver, por un lado, con qué problema se quiere solucionar con una solución algorítmica y, por otro, con los objetivos que pueden esperarse del modelo. A menudo, las visiones reduccionistas sobre estas fases, o bien infraestiman la dificultad de definir el problema y su contexto, o bien sobreestiman la capacidad de las funciones algorítmicas para poner solución a problemas complejos. Todo ello puede generar distorsiones con un gran impacto en el despliegue del modelo[160]; un impacto a menudo imprevisible por operadores que intervienen en fases posteriores, que no son conscientes de las problemáticas subyacentes a un desarrollo o despliegue aparentemente correctos.

La AEPD denomina a esta fase "concepción y análisis", en la que se fijan los requisitos funcionales y no funcionales del sistema, que *vendrán fijados por objetivos de negocio derivados del tratamiento en donde se incorporará o del mercado donde se pretende comercializar el componente* (lo cual incluye los planes de proyecto, las restricciones normativas, etc)[161]. En el ámbito de Derecho Público, las decisiones que se adoptan en esta fase tienen un *alto* contenido político[162]. Así, no puede obviarse que la propia decisión de automatizar determinado proceso o procedimiento administrativo es una decisión fundamentalmente política en dicho contexto. Además, Palma recuerda que la colaboración público-privada es cada vez será más frecuente para el desarrollo de sistema de decisiones automatizadas, *en donde las distintas organizaciones aportan diversos elementos que ayuden a conformar el sistema y donde ambas se beneficiarán del producto creado*[163].

Desde un punto de vista ético, las decisiones a adoptar en esta fase nos empujan a reflexionar sobre los beneficios y perjuicios de datificar, automatizar y regular de forma

[160] Este impacto puede tener un carácter económico – por ejemplo, la pérdida de una costosa inversión por el desarrollo de un sistema incapaz de ofrecer soluciones satisfactorias en el mundo real-, ético -por el desarrollo de un sistema que manipule la conducta de sujetos vulnerables- o jurídico -por el desarrollo de un sistema que viola derechos fundamentales-.

[161] Agencia Española de Protección de Datos (AEPD), «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción», 12.

[162] de Fine Licht y de Fine Licht, «Artificial intelligence, transparency, and public decision-making», 4.

[163] Palma Ortigosa, «Régimen jurídico de la toma de decisiones automatizadas y el uso de sistemas de inteligencia artificial en el marco del derecho a la protección de datos personales», 60.

algorítmica un proceso social determinado, en definitiva, sobre si es posible y deseable dar una solución automatizada a determinado problema. E incluso sobre si todo lo que es técnicamente posible, resulta éticamente aceptable. Como ejemplo de ello, la comunidad internacional debate sobre el desarrollo de los Sistemas de Armas Autónomos Letales (SAAL). Aunque no faltan voces a favor del diseño y desarrollo de los SAAL[164], António Guterres, noveno Secretario General de las Naciones Unidas, ha manifestado públicamente su rechazo político y moral a estos sistemas que, a su modo de ver, deben prohibirse por el Derecho Internacional.

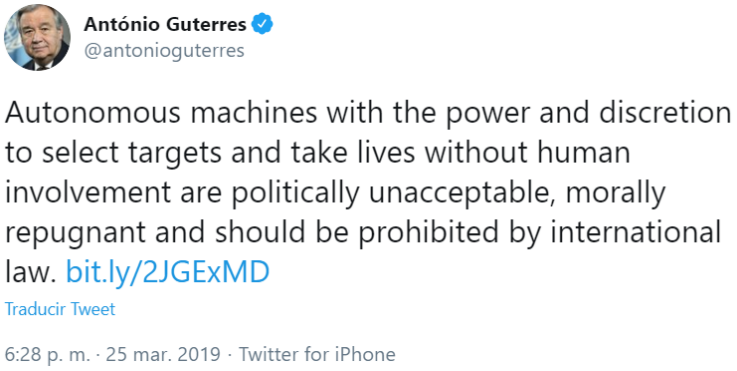

Ilustración 1. Tweet de @antonioguterres

En definitiva, desde un punto de vista normativo, estas reflexiones pueden traducirse en prohibiciones para el desarrollo de determinados modelos para su aplicación en ámbitos concretos, o al menos en limitaciones al diseño de los mismos. En esta línea, la Comisión Europea ha propuesto la creación de una categoría de sistemas de IA de riesgo inaceptable por ser contrario a los valores de la Unión -de respeto de la dignidad humana, libertad, igualdad, democracia y Estado de Derecho y de los derechos fundamentales- y que deben ser prohibidas. Las prohibiciones engloban aquellas prácticas con potencial para manipular a las personas o para alterar de manera sustancial su comportamiento de un modo que es probable que les provoque perjuicios físicos, así como que las autoridades públicas realicen calificación social basada en IA con fines generales o, salvo excepciones limitadas, el uso de sistemas de identificación biométrica remota en tiempo real en espacios de acceso público con fines de aplicación de la ley[165].

[164] Un resumen de este debate puede encontrarse en Rubio Damián, «Automatización de la guerra: el control humano».

[165] La lista completa se ubica actualmente en el artículo 5 de la propuesta AIA.

1.1.2. Desarrollo: recolección de datos, entrenamiento y validación del modelo

La fase de desarrollo consiste en que el modelo alcance los estándares fijados en la fase de diseño[166]. Este proceso incluye una multitud de etapas que han sido simplificadas en este apartado y que son comunes al desarrollo de los modelos de aprendizaje automático que más interés reúnen en la actualidad: recolección de datos, entrenamiento y validación -aunque no siempre han de estar presente en toda solución algorítmica[167].

En esta fase de desarrollo se produce la construcción de la "experiencia de la máquina" para la posterior implementación de esa experiencia en el mundo real. El primer escalón en la construcción de esa experiencia se relaciona con la producción y disponibilidad de datos fiables y precisos que sean adecuados para entrenar un modelo algorítmico[168].

La clase de datos, de carácter personal o no, que se van a utilizar para el entrenamiento y validación del modelo tiene una relevancia jurídica fundamental. Si estos son de carácter personal, se habrá de aplicar el RGPD y será necesario el establecimiento de una base legitimadora, tanto para la recolección de los datos como para el entrenamiento y validación de los modelos -bases que varían, a su vez, en función de si se incluyen datos de carácter especial o no-[169]. Al contrario, si el entrenamiento se va a realizar con datos que no tienen carácter personal, no será de aplicación el RGPD[170]. Ahora bien, merece la

---

[166] de Fine Licht y de Fine Licht, «Artificial intelligence, transparency, and public decision-making», 4.

[167] Agencia Española de Protección de Datos (AEPD), «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción», 12.

[168] Vid. Cabitza, Campagner, y Balsano, «Bridging the "Last Mile" Gap between AI Implementation and Operation: "Data Awareness" That Matters».

[169] Por supuesto la aplicación del RGPD no se limita a la existencia de una base legitimadora del tratamiento, para un examen más exhaustivo vid. Agencia Española de Protección de Datos (AEPD), «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción», 20 y ss.

[170] Considerando 26 RGPD sobre la aplicación del mismo a datos seudonimizados -datos personales- y no a anonimizados: *Los principios de la protección de datos deben aplicarse a toda la información relativa a una persona física identificada o identificable. Los datos personales seudonimizados, que cabría atribuir a una persona física mediante la utilización de información adicional, deben considerarse información sobre una persona física identificable. Para determinar si una persona física es identificable, deben tenerse en cuenta todos los medios, como la singularización, que razonablemente pueda utilizar el responsable del tratamiento o cualquier otra persona para identificar directa o indirectamente a la persona física. Para determinar si existe una probabilidad razonable de que se utilicen medios para identificar a una persona física, deben tenerse en cuenta todos los factores objetivos, como los costes y el tiempo necesarios para la identificación, teniendo en cuenta tanto la tecnología disponible en el momento del tratamiento como los avances tecnológicos. Por lo tanto, los principios de protección de datos no deben aplicarse a la información anónima, es decir información que no guarda relación con una persona física identificada o identificable, ni a los datos convertidos en anónimos de forma que el interesado no sea identificable, o deje de serlo. En consecuencia, el presente Reglamento no afecta al tratamiento de dicha información anónima, inclusive con fines estadísticos o de investigación*.

pena recalcar que el proceso de anonimización de datos personales es un tratamiento de datos personales que entra en el ámbito del RGPD[171]. Una vez anonimizados -o sintetizados[172]-, esto es, dichos datos no tienen carácter personal.

Tampoco está de más recalcar que los algoritmos de aprendizaje automático son algoritmos *hambrientos* de datos, a diferencia del aprendizaje humano que puede basarse en la capacidad de representar relaciones abstractas a partir de pocos ejemplos, el rendimiento de estos algoritmos se basa en miles, millones o incluso miles de millones de ejemplos de entrenamiento[173]. En este sentido, el desarrollo de estos modelos requiere un intenso "trabajo de datos" -cualquier actividad humana relacionada con la creación, recolección, gestión, limpieza, análisis, interpretación y comunicación de datos[174]-, tal es la magnitud de estas tareas que únicamente la limpieza, reorganización y preprocesamiento de los datos consumen *la mayor parte* de un proyecto de IA[175].

Es decir, entre la recolección de los datos y el entrenamiento de un modelo algorítmico existe un paso intermedio que requiere de una cantidad de recursos a menudo infraestimada, donde afloran problemáticas sobre la interoperabilidad y reproducibilidad

[171][171] Acerca de las distintas técnicas de anonimización, vid. Agencia Española de Protección de Datos (AEPD), «La K-Anonimidad como medida de la privacidad». Disponible en: https://www.aepd.es/sites/default/files/2019-09/nota-tecnica-kanonimidad.pdf

[172] Los datos sintéticos son aquéllos que, a partir de una base de datos, son generados de forma automatizada simulando los datos de origen manteniendo las correlaciones estadísticas que contenía la base de datos original. Esta clase de datos son utilizados, en definitiva, para disociar la información original de las bases originales, mientras que se mantiene su valor estadístico para ser utilizados, por ejemplo, en el entrenamiento de modelos algorítmicos. En términos jurídicos no hay ninguna consideración específica para los datos sintéticos, esto implica que los datos sintéticos habrán de ser o datos personales o datos anónimos. Esto es, según las características de la técnica de sintetización de la base de datos original, podremos considerar el resultado de dicha técnica como datos personales -seudonimizados- o anonimizados y aplicar, en consecuencia, el régimen jurídico pertinente a cada cual. No obstante, no puede olvidarse una vez más que, para la generación de estos datos, será necesaria la legitimación para el acceso a los datos personales y para su tratamiento de sintetización.

[173] Marcus, «Deep Learning: A Critical Appraisal». Aunque esto no es un requisito necesario para el desarrollo de la IA: *Different technologies require different resources. While AI is sometimes associated with big data applications that rely on training or analysis of huge amounts of data, big data is not a necessary requirement. There are also small data applications or applications that do not require significant training data at all. The resources vary accordingly*. Djeffal, «AI, Democracy and the Law», 257-58.

[174] Bossen et al., «Data work in healthcare: An Introduction», 466.

[175] Núñez Reiz, Armengol de la Hoz, y Sánchez García, «Big Data Analysis y Machine Learning en medicina intensiva». Estos autores explican que la limpieza, reorganización y preprocesamiento consiste en, una vez extraídos los datos verificar su calidad y darles el formato y escala adecuados, resolviendo problemas de ausencias e inconsistencias para prepararlos para ser procesados por el modelo correspondiente.

de los datos más allá de su disponibilidad para la recolección[176]. Con carácter general, estas tareas previas al entrenamiento se conocen como actividades de preprocesamiento, y tienen un impacto considerable en la probabilidad de cumplir con los objetivos fijados en el diseño del sistema.

Este impacto del preprocesamiento es considerable tanto a efectos de rendimiento del modelo -problemas de sobreajuste u *overfitting*-[177], como para la reproducción de sesgos que pudieran dar lugar a la aparición de resultados discriminatorios en la implementación del sistema[178]. El preprocesamiento es determinante a la hora de establecer una representatividad apropiada en los conjuntos de datos, puesto que en la recolección de datos no es habitual que esta representatividad venga dada, lo que provoca distorsiones en los resultados algorítmicos tanto si se da una infrarrepresentación como una sobrerrepresentación de algún grupo poblacional, con independencia de la "calidad" de los registros individuales[179]. De ahí, la relevancia normativa que adquiere el preprocesamiento junto a la recolección de datos, tanto en el análisis de la legislación

[176] A nivel normativo, Sánchez Caro señala la falta de interoperabilidad como uno de los mayores obstáculos para la digitalización de la salud, y dice en particular que la ausencia de normas que obliguen a la interoperabilidad impide la innovación y limita la utilización en escala de las soluciones en las que invertimos, vid. Sanchez Caro, «Cambio de paradigma en la relación clínico-asistencial: Aspectos bioéticos y legales». En este sentido, la Directiva 2011/24/UE, relativa a la aplicación de los derechos de los pacientes en la asistencia sanitaria transfronteriza, ilustra una estrategia normativa de cooperación voluntaria no ha surtido los efectos deseados en su Considerando 56: *La presente Directiva debe reconocer por ello tanto la importancia de trabajar en favor de la interoperabilidad, por una parte, como la adecuada división de competencias, por otra, disponiendo a tal fin lo necesario para que la Comisión y los Estados miembros sigan cooperando en la elaboración de medidas que, sin ser jurídicamente vinculantes, constituyan herramientas entre las que los Estados miembros puedan elegir para facilitar una mayor interoperabilidad de los sistemas de tecnologías de la información*. Una de las iniciativas que pretende revertir esta insuficiente estrategia para favorecer la interoperabilidad de los datos en salud y otros ámbitos de aplicación es la creación de espacios europeos de datos propuesta por la Estrategia Europea de Datos de la Comisión Europea.

[177] Acerca del fenómeno del *overfitting*, se recomienda la explicación de Palma en lengua castellana. Vid. Palma Ortigosa, «Régimen jurídico de la toma de decisiones automatizadas y el uso de sistemas de inteligencia artificial en el marco del derecho a la protección de datos personales», 76-77.

[178] Sobre esto último, más ampliamente el apartado 3. Sesgos en la toma de decisiones automatizada basada en la elaboración de perfiles, en este mismo capítulo.

[179] Barocas y Selbst, «Big Data's Disparate Impact», 684-86.

vigente[180], como en las iniciativas para la regulación de los sistemas de IA de alto riesgo[181].

La fase de entrenamiento es, en realidad, la fase en la cual se construye un determinado modelo algorítmico a partir de un conjunto de datos determinado -previamente recolectado y preprocesado-, de forma que dicho modelo podrá posteriormente realizar predicciones sobre objetos futuros -que no forman parte de ese conjunto de datos determinado-.

Dicha fase exige elegir una clase de algoritmo que sirva de la mejor manera posible a los objetivos dados para el sistema, lo cual requiere, a su vez, escoger entre diferentes funciones (entre otras, clasificación -el resultado es una clase, entre un número limitado de clases como "spam" o "no spam"- o regresión -cuyo objetivo es predecir valores continuos, como el precio estimado de un inmueble-), diferentes técnicas de aprendizaje (supervisado, no supervisado o por refuerzo). En palabras de Palma, la elección de uno u otro algoritmo va a estar estrechamente vinculada al tipo de problema que pretenda resolver la organización con el proyecto que despliega[182]. En su propuesta de Reglamento AIA, la Comisión opta por calificar como sistemas de IA una amplia variedad de algoritmos[183].

---

[180] En este sentido, la AEPD: *El hecho de alimentar un modelo de aprendizaje de IA con datos sin ningún control ni análisis previo, además de no estar justificado, especialmente en el caso de que el tratamiento se base en el interés legítimo, puede hacer que la IA pierda precisión y se convierta en un multiplicador de sesgos. El conjunto de datos a utilizar ha de analizarse cuidadosamente para evitar dichos riesgos y legitimar su uso si, por ejemplo, el tratamiento se está basando en el interés legítimo*. Vid. Agencia Española de Protección de Datos (AEPD), «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción».

[181] De ahí la relevancia que en la propuesta de Reglamento AIA de la Comisión ha adquirido la gobernanza de datos como requisito obligatorio para los sistemas de IA de alto riesgo, estableciendo obligaciones concretas para la recolección y el preprocesamiento para los proveedores de sistemas. Así se expresa en el Considerando 44: Es preciso instaurar prácticas adecuadas de gestión y gobernanza de datos para lograr que los conjuntos de datos de entrenamiento, validación y prueba sean de buena calidad. Los conjuntos de datos de entrenamiento, validación y prueba deben ser lo suficientemente pertinentes y representativos, carecer de errores y ser completos en vista de la finalidad prevista del sistema.

[182] Palma Ortigosa, «Régimen jurídico de la toma de decisiones automatizadas y el uso de sistemas de inteligencia artificial en el marco del derecho a la protección de datos personales», 81.

[183] Anexo I AIA. Técnicas y estrategias de IA: *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*.

Ahora bien, la elección de un determinado algoritmo frente a otro no es una mera cuestión de ajustar mejor o peor el rendimiento estadístico del mismo a los problemas fijados. Esta elección tiene una gran repercusión sobre la interpretabilidad de los modelos y, por ende, sobre la capacidad para comprender el peso de los distintos atributos y características en los resultados del modelo, lo cual tiene a su vez influencia sobre la posible reproducción y mitigación de sesgos algorítmicos y sobre las posibilidades de supervisión humana en la fase de implementación[184]. Tampoco puede pasarse por alto que algunos modelos pueden aprender de forma constante y una vez implementados, es decir, el entrenamiento en estos modelos puede continuar produciéndose en la fase de implementación[185].

Una vez entrenado un determinado modelo, éste será capaz de elaborar inferencias de forma automatizada a partir de nuevos datos de entrada que se le presenten. Para poder testar si el modelo entrenado responde de forma satisfactoria ha de llevarse a cabo un proceso de validación. En definitiva, en este contexto "validar" significa aportar evidencias de que el modelo es sólido, es decir, que funcionará correctamente con nuevos datos que el modelo nunca ha examinado o procesado antes.

Los resultados de estas evidencias pueden expresarse en distintas métricas, habitualmente expresadas a través de la matriz de confusión, que permite evaluar el rendimiento de un modelo a partir de la clase de aciertos y errores que comete[186]. La evaluación de estas distintas métricas -exactitud, precisión, sensibilidad y especificidad-, tan habitual por ejemplo en la expresión del rendimiento de pruebas diagnósticas en el ámbito clínico,

[184] Sobre todas estas cuestiones se ampliará en los apartados correspondientes.

[185] En el plano normativo pueden establecerse obligaciones particulares para los sistemas que aprenden de forma continua, así lo recoge la propuesta de Reglamento AIA en sus artículos 15(3) y 43(4). Considerando 66 AIA*: En consonancia con la noción comúnmente establecida de «modificación sustancial» de los productos regulados por la legislación de armonización de la Unión, conviene que un sistema de IA se someta a una nueva evaluación de la conformidad cada vez que se produzca un cambio que pueda afectar al cumplimiento por su parte del presente Reglamento o cuando la finalidad prevista del sistema cambie. Por otro lado, 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.*

[186] Palma realiza una excelente exposición sobre la matriz de confusión y de las métricas de evaluación del rendimiento del modelo. Vid. Palma Ortigosa, «Régimen jurídico de la toma de decisiones automatizadas y el uso de sistemas de inteligencia artificial en el marco del derecho a la protección de datos personales», 89-99.

define su relevancia en función del contexto, dado que la importancia ético-jurídica de un falso positivo o de un falso negativo tiene un valor muy distinto en contextos diversos[187].

En cuanto a las formas de validación, podría distinguirse entre validación interna y externa[188]. El objetivo de la validación interna es evaluar el rendimiento predictivo de un modelo en datos no vistos con respecto al entrenamiento del modelo, pero que provienen de la misma población y entorno[189]. Es decir, se refiere a los protocolos de validación que intentan estimar el rendimiento de los modelos dividiendo el conjunto de datos de entrenamiento en múltiples conjuntos de datos más pequeños, y probando el modelo -entrenado con una parte del conjunto de datos original- en una parte diferente, normalmente más pequeña[190]. De esta forma, se trata de evaluar si el modelo, a partir de las correlaciones aprendidas en la fase de entrenamiento, es capaz de generalizar de forma adecuada sobre datos nuevos, pero con una representatividad similar a la de los datos de entrenamiento[191]. Se trata, en definitiva, de una etapa necesaria en el desarrollo de un modelo para comprobar que el modelo entrenado funciona satisfactoriamente, pero que arroja poca información acerca de cuál será el rendimiento del modelo en la fase de implementación.

Con validación externa, por el contrario, nos referimos a la realizada con conjuntos de datos que provienen de cohortes o repositorios distintos de los conjuntos de datos

---

[187] Por ejemplo, Urruela Mora en relación con el uso de sistemas de evaluación del riesgo en el ámbito penal indica que el valor relevante a efectos de la estimación a efectuar en el ámbito del sistema de justicia penal es fundamentalmente el valor predictivo. Y advierte del riesgo que puede generar la presentación de los propios resultados y de los estudios de validación de la herramienta de predicción del riesgo a profanos -incluidos juristas-, si personas especializadas no discriminan adecuadamente sobre dichos parámetros. Vid. Urruela Mora, «Instrumentos de evaluación del riesgo de violencia, justicia algorítmica y derecho penal. Perspectiva crítica».

[188] La Comisión en su propuesta de Reglamento AIA distingue entre validación, equivalente a validación interna en esta investigación, y prueba, equivalente a validación externa previa a la comercialización. Los define en el artículo 3 así: «Datos de validación»: *los datos usados para proporcionar una evaluación del sistema de IA entrenado y adaptar sus parámetros no entrenables y su proceso de aprendizaje, entre otras cosas, para evitar el sobreajuste. El conjunto de datos de validación puede ser un conjunto de datos independiente o formar parte del conjunto de datos de entrenamiento, ya sea como una división fija o variable.* (…) «Datos de prueba»: *los datos usados para proporcionar una evaluación independiente del sistema de IA entrenado y validado, con el fin de confirmar el funcionamiento previsto de dicho sistema antes de su introducción en el mercado o su puesta en servicio.*

[189] de Hond et al., «Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review», 4.

[190] Cabitza et al., «The importance of being external. methodological insights for the external validation of machine learning models in medicine», 1-2.

[191] Lo habitual y exigible es que la exactitud de estos modelos en la validación interna sea muy alta (por encima de un 90-95%) para poder considerar que el entrenamiento del modelo es satisfactorio.

utilizados para la creación del modelo[192]. La AEPD señala que en esta operación se podría realizar un tratamiento de datos personales cuando se utilicen datos que corresponden a la situación real del tratamiento, para determinar la solidez del modelo de forma experimental, pudiendo realizarse por un tercero para la auditoría o certificación del modelo[193]. La mayoría de las veces, el rendimiento observado en los conjuntos de datos externos es significativamente inferior al rendimiento evaluado en los conjuntos de datos originales[194].

La validación externa se realiza habitualmente durante la fase de desarrollo previa a la comercialización del modelo, de forma que la normativa puede establecer los estándares o procedimientos de validación necesarios que determinen la aceptabilidad de un sistema, pudiendo imponer un sistema de autovalidación -realizada bajo responsabilidad del propio desarrollador del sistema- o de heterovalidación -cuando se realiza bajo responsabilidad de terceros-. El modelo regulatorio que la Comisión propone para los sistemas de IA de alto riesgo consiste en un sistema de gestión de riesgos que incluye la realización de pruebas -validación externa- adecuadas a la finalidad prevista del sistema, que ha de documentarse y someterse a evaluación para obtener la certificación oportuna[195].

No obstante, la validación externa puede continuar produciéndose una vez que el sistema está comercializado, es decir, en su fase de implementación. Es habitual que la regulación establezca requerimientos para la realización de seguimiento posterior a la comercialización que incluyan la validación continua del modelo. Aunque no incluye sistemas de validación concretos, la propuesta de Reglamento AIA de la Comisión

[192] de Hond et al., «Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review», 6.

[193] Agencia Española de Protección de Datos (AEPD), «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción», 12-13.

[194] Cabitza et al., «The importance of being external. methodological insights for the external validation of machine learning models in medicine», 1-2.

[195] Art. 9 AIA: 5. *Los sistemas de IA de alto riesgo serán sometidos a pruebas destinadas a determinar cuáles son las medidas de gestión de riesgos más adecuadas. Dichas pruebas comprobarán que los sistemas de IA de alto riesgo funcionan de un modo adecuado para su finalidad prevista y cumplen los requisitos establecidos en el presente capítulo; 6. Los procedimientos de prueba serán adecuados para alcanzar la finalidad prevista del sistema de IA y no excederán de lo necesario para ello; 7. Las pruebas de los sistemas de IA de alto riesgo se realizarán, según proceda, en cualquier momento del proceso de desarrollo y, en todo caso, antes de su introducción en el mercado o puesta en servicio. Los ensayos se realizarán a partir de parámetros y umbrales de probabilidades previamente definidos que sean adecuados para la finalidad prevista del sistema de IA de alto riesgo de que se trate*.

impone un seguimiento durante la implementación de los sistemas de IA[196] -similar al modelo de gestión de riesgo conforme al principio de responsabilidad en el RGPD[197]-, aunque menos estrictos de los que se imponen, por ejemplo, en la normativa europea sobre productos sanitarios o de productos in vitro.

1.2. Implementación del modelo

Con implementación del modelo, se hace referencia a su utilización en el entorno social -vía comercial o no- produciendo inferencias a partir de datos de entrada nunca antes procesados por el mismo. A nivel normativo, puede imponerse determinada certificación o evaluación previa a la implementación del modelo, si no fuere así, el propio desarrollador o proveedor del sistema decidirá el momento oportuno en el que, tras realizar el entrenamiento y validación oportunas, decide implementar el modelo.

Durante la implementación, el modelo funcionará generando datos de salida -*outputs*- realizando correlaciones de forma autónoma a partir del aprendizaje concreto del modelo con los datos de entrada -*inputs*- que reciba. Con carácter general, podemos decir que esos datos de salida son inferencias que realiza el modelo[198], pudiendo tratarse de una clasificación, análisis o predicción sobre el entorno que le rodea y percibe a través de los datos de entrada.

A los efectos de esta investigación, resultan relevantes las inferencias que se refieren a una persona física y sus aspectos personales, lo que en términos de la normativa de protección de datos se conoce como elaboración de perfiles[199]. Ahora bien, no todos los

---

[196] Art. 3 AIA: (…) «Seguimiento posterior a la comercialización»*: todas las actividades realizadas por los proveedores de sistemas de IA destinadas a recopilar y examinar de forma proactiva 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*.

[197] Art. 24 RGPD: *1. Teniendo en cuenta la naturaleza, el ámbito, el contexto y los fines del tratamiento así como los riesgos de diversa probabilidad y gravedad para los derechos y libertades de las personas físicas, el responsable del tratamiento aplicará medidas técnicas y organizativas apropiadas a fin de garantizar y poder demostrar que el tratamiento es conforme con el presente Reglamento. Dichas medidas se revisarán y actualizarán cuando sea necesario.*

[198] Diccionario de la lengua española (RAE), "inferir": *Del lat. inferre 'llevar a'. 1. tr. Deducir algo o sacarlo como conclusión de otra cosa*.

[199] Art. 4 RGPD: 4) «elaboración de perfiles»: *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*. Korff expone que estos perfiles anteriormente eran bastante simples, por ejemplo,

modelos no van a realizar inferencias necesariamente personales, esto es, perfiles; por ejemplo, un modelo para conducción autónoma elaborará inferencias para que el vehículo transite sin realizar perfiles personales -o al menos sin que éstos sean la base del funcionamiento del modelo-. Del mismo modo, el carácter personal o no de los datos de entrada tiene consecuencias en la aplicación de la normativa de protección de datos, dado que la utilización de dichos datos para generar perfiles se considera un tratamiento de datos[200].

A nivel normativo no solo es relevante si la inferencia es o no de carácter personal, sino también el fin de la propia inferencia, por ejemplo, en el ámbito de la salud es distinto un sistema cuyas inferencias tengan por objeto evaluar el ritmo cardíaco de una persona para mejorar el rendimiento deportivo, que evaluar el ritmo cardíaco durante una cirugía para determinar el riesgo de paro cardíaco de un paciente[201].

Siguiendo lo que recoge la AEPD -aunque lo haga desde la perspectiva de la protección de datos-, quien decide adoptar en fase de implementación una solución técnica basada en IA, o en cualquier otra tecnología, es responsable en sentido amplio de las consecuencias éticas y jurídicas de incorporar esta herramienta a la toma de decisiones. Así lo explica: «*la decisión de adoptar, en el marco de un tratamiento, una solución técnica basada en IA o en cualquier otra tecnología, es tomada por el responsable, que es quien "determina los medios y fines del tratamiento" y es, por tanto, tiene a su cargo la toma de decisión de seleccionar una solución tecnológica u otra. En dicho responsable descansa la obligación de ser diligente a la hora de seleccionar la más adecuada, en particular cuando contrata su desarrollo o la adquiere; exigir y analizar las especificaciones de calidad de la solución; y determinar la extensión del tratamiento y la*

en publicidad podía considerarse un perfil de hombres de entre 20 y 25 años con una renta disponible de más de veinte mil euros al año podían considerarse lectores "típicos" de una determinada revista. Lo novedoso de los perfiles actuales basados en los modelos descritos es el análisis estadístico y la ponderación. Esencialmente, los perfiles modernos (como los antiguos) consisten en una lista de factores vinculados a una cuestión o resultado concreto, pero con un peso determinado, posiblemente dinámico, para cada factor. Ello tiene como consecuencia la elaboración de unos perfiles mucho más específicos y personalizados o, mejor dicho, atomizados. Vid. Korff, «Comments on Selected Topics in the Draft EU Data Protection Regulation», 51.

[200] No obstante, a efectos de la presente investigación es irrelevante que los perfiles hayan sido elaborados total o parcialmente a partir de datos personales.

[201] El segundo ejemplo tiene un "fin médico específico" conforme al artículo 2.1 del Reglamento (UE) 2017/745 sobre Productos Sanitarios, mientras que el primero no. Por ende, dado el fin de la inferencia del segundo ejemplo, será aplicable al diseño y desarrollo del segundo las obligaciones de dicho Reglamento para su certificación y posterior comercialización e implementación.

carga de hacer frente a las consecuencias de sus decisiones. El que toma la decisión de realizar el tratamiento es responsable, y no puede escudarse en la carencia de información o el desconocimiento técnico para evadir su responsabilidad a la hora de auditar y decidir la adecuación del sistema»[202].

En la fase de implementación, puede distinguirse a su vez entre la "inferencia" y la "decisión". Ello depende, por supuesto, de la definición normativa que se adopte por decisión. A los efectos de esta investigación, se entiende por decisión la producción de un efecto jurídicamente relevante sobre la persona que se realiza la inferencia. Este efecto puede producirse directamente por la propia realización de la inferencia, porque despliega directamente este efecto, o únicamente tras la aplicación de esta inferencia con intervención humana o no[203]. La relevancia jurídica puede darse tanto por desplegar un efecto normativamente previsto -la decisión da lugar a una investigación tributaria normativamente prevista-, como por desplegar un efecto sobre un bien jurídico protegido en el ámbito normativo -decisión que afecta a la salud-.

Hay razones de peso para restringir el peso de las inferencias -que reflejan necesariamente un estado de las cosas pasado- sobre las decisiones que nos afectan: «*nos equivocamos si confiamos ciegamente en un mecanismo cuyos resultados son un fiel reflejo de las aportaciones del pasado que lo han configurado. Por lo tanto, debemos restringir las conclusiones que sacamos de esos resultados y las acciones que emprendemos, o de lo contrario estaremos equivocándonos en el futuro*»[204]. Es por ello que la restricción normativa sobre la aplicabilidad directa de las inferencias es un tema central en la regulación de los sistemas algorítmicos. Esta restricción adopta habitualmente la forma de participación humana obligatoria en la fase de implementación, es decir, se impone una clase de intervención/supervisión por un agente humano, previa o simultánea a la producción de efectos de la inferencia.

---

[202] Agencia Española de Protección de Datos (AEPD), «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción», 19.

[203] La normativa de protección de datos distingue efectivamente entre la figura del perfil -art. 4(4) RGPD- y la decisión propiamente dicha -a efectos del art. 22 RGPD-, lo cual no implica que la elaboración de un perfil no pueda constituir por sí una decisión automatizada. Vid. Apartado 3.1. Ubicación de la toma de decisiones automatizada y de la elaboración de perfiles en el RGPD. Capítulo 2. Toma de decisiones automatizada basada en la elaboración de perfiles en el RGPD: el artículo 22 en la unidad de cuidados intensivos.

[204] Grant y Wischik, «Poisonous Datasets, Poisonous Trees», 95-96.

Así, la implementación de un modelo algorítmico puede producirse en distintos "niveles" de mayor a menor automatización[205]. Este grado de automatización se relaciona directamente con la clase de interacción humana que se imponga en la implementación del modelo y de la relación de esta interacción con las inferencias que se defina, determinando por ejemplo los fines o límites de dicha interacción. A su vez, desde el ordenamiento jurídico existen numerosas formas de construir el significado de las inferencias algorítmicas en función del grado de automatización o, dicho de otra forma, del papel que se otorgue a interacción humana con el sistema. En este sentido, las inferencias pueden convertirse en ilegales e irrelevantes, pueden someterse a la supervisión y a la toma de decisiones humanas, o incluso se les puede otorgar directamente fuerza de ley[206].

Por último, es oportuno recordar dos procesos que pueden ser habituales en esta fase y ocurren también en anteriores fases. Por un lado, el entrenamiento continuo de los sistemas que siguen aprendiendo después de su introducción en el mercado o puesta en servicio durante su implementación, es decir, aquellos que adaptan automáticamente el modo en que desempeñan sus funciones[207]. Y, por otro lado, los procesos de validación del modelo durante la fase de implementación, también mencionados anteriormente.

Esta clase de validación, que Cobbe, Lee y Singh denominan fase de investigación[208], puede adoptar distintas formas en el plano normativo. Al igual que durante el diseño y desarrollo, puede tratarse de un proceso de autovalidación, como el adoptado por el

[205] Burns, «Where is the evidence for automated triage apps?», 20.Siempre resulta útil recordar en este punto los niveles de automatización de la conducción definidos en el informe del SAE, que van desde la no automatización de la conducción (nivel 0) hasta la automatización completa de la misma (nivel 5). Vid. Society of Automotive Engineers (SAE), «Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_201806».

[206] Djeffal, «AI, Democracy and the Law», 273. Esta cuestión se abordará de forma específica más adelante. Vid. Apartado 2. La intervención humana en la toma de decisiones automatizada basada en la elaboración de perfiles, en este mismo capítulo.

[207] La guía de la AEPD habla de la "evolución" de los sistemas en base a datos personales: en la solución IA se podrían usar los datos y resultados de los interesados para refinar el modelo de IA. Cuando nos encontramos que esa evolución se realiza en el componente adquirido por el propio interesado, de forma aislada y autónoma, aplicaría la excepción doméstica. Pero si se envían a terceros, tendríamos una comunicación de datos, un posible tratamiento de almacenamiento, tratamiento para modificar el modelo, o incluso nuevas comunicaciones si esos datos se incorporan al modelo y este es accesible a otros terceros Agencia Española de Protección de Datos (AEPD), «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción», 13.

[208] Cobbe, Lee, y Singh, «Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems», 9.

RGPD para tratamientos de datos de alto riesgo, para los que es obligatoria la realización de una Evaluación de Impacto de Protección de Datos (EIPD). La EIPD debe realizarse antes del tratamiento (art. 35(1) y (10) RGPD), pero ello no quiere decir que no deba actualizarse, realizándose la evaluación de forma continua mientras persista el tratamiento de datos: «*La actualización de la EIPD a lo largo del proyecto de ciclo de vida garantizará que se tenga en cuenta la protección de los datos y la intimidad y propiciará la creación de soluciones que fomenten el cumplimiento*[209]. Del mismo modo, la evaluación del funcionamiento del modelo puede darse por agentes externos, que habitualmente pueden ser reguladores o entidades autorizadas por la regulación para llevar a cabo tareas de auditoría»[210].

En definitiva, las fases aquí descritas permiten visualizar de forma atomizada la toma de decisiones automatizada basada en la elaboración de perfiles y la relevancia jurídica que adquieren la multitud de decisiones adoptadas en este proceso. Este ejercicio descriptivo no tenía la pretensión, ni mucho menos, de ser exhaustivo. No obstante, esta visión general del proceso permitirá entender las cuestiones particulares que se abordan -ahora sí con mayor exhaustividad- a continuación.

2. La intervención humana en la toma de decisiones automatizada basada en la elaboración de perfiles

Según Favaretto et al. las tecnologías de *big data* están indisolublemente ligadas a una dicotomía en virtud de la cual los seres humanos son tanto la causa de sus defectos, como los supervisores de su correcto funcionamiento[211].

Tal y como esta investigación recoge en el apartado referido a la implementación del modelo algorítmico en el entorno real, dicha implementación puede producirse con distintos "niveles" de mayor a menor automatización[212]. Sin embargo, estas clasificaciones no son el resultado de un aumento "cuantitativo" de la intervención

[209] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «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», 16.

[210] Cobbe, Lee, y Singh, «Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems», 9.

[211] Favaretto, De Clercq, y Elger, «Big Data and discrimination: perils, promises and solutions. A systematic review», 21.

[212] Burns, «Where is the evidence for automated triage apps?», 20.

humana, sino más bien una forma de calificar la forma en que la agencia y el control humanos operan en un contexto automatizado. La interacción o participación humana está presente y tiene sus efectos sobre todas las fases de la toma de decisiones automatizada, no obstante, encontramos un interés particular en la introducción de esta interacción durante la implementación del sistema para servir a determinados objetivos normativos[213]. Ahora, ello no quiere decir que dicha intervención sea necesariamente más garantista, en el sentido de servir mejor a los objetivos normativos que se establezcan, por producirse en la fase de implementación o despliegue del sistema[214].

A su vez, desde el ordenamiento jurídico existen numerosas formas de construir el significado de los resultados de los modelos algorítmicos en función del papel que se otorgue a la intervención humana. Así, puede abrirse la construcción social de la tecnología a la deliberación y la toma de decisiones democráticas; los resultados de los sistemas de IA pueden convertirse en ilegales o irrelevantes, pueden someterse a la supervisión y a la toma de decisiones humanas, o incluso se les puede otorgar directamente fuerza de ley[215].

Sobre las múltiples formas -supervisión, intervención, control- que puede adoptar la participación humana en un contexto normativo, las posibilidades son amplias teniendo en cuenta que dicha interacción o participación puede responder a distintos principios u objetivos normativos. También la distinta cualificación que este requerimiento normativo puede adoptar -significativo, adecuado, efectivo-. Sirva de muestra gráfica la siguiente tabla elaborada en el seno del debate del Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales de la Convención sobre prohibiciones o restricciones del empleo de ciertas armas convencionales que puedan considerarse excesivamente nocivas o de efectos indiscriminados[216]:

[213] A esta clase de interacción o participación humana de carácter normativo en la fase de implementación y despliegue de un sistema, la denominaré con carácter general "intervención humana".

[214] Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems», 3.

[215] Djeffal, «AI, Democracy and the Law», 273. Menciona en este sentido el papel del artículo 22 RGPD, que se analizará en profundidad en el siguiente capítulo, y cataloga el mismo como un derecho de supervisión humana que somete las decisiones totalmente automatizadas a las decisiones humanas.

[216] Esta convención fue ratificada por España en «BOE» núm. 89, de 14 de abril de 1994, páginas 11384 a 11393.

(Mantener) (Garantizar) (Ejercer) (Conservar)	un nivel	(sustantivo) (significativo) (apropiado) (suficiente) (mínimo) (mínimo indispensable)	de	(participación) (implicación) (responsabilidad) (supervisión) (validación) (control) (discernimiento) (decisión)	por parte del ser humano

Tabla 2. CCW/GGE.1/2018/3[217]

En este apartado, primero se aborda el acercamiento técnico a la intervención humana en procesos de toma de decisiones automatizada a partir del uso de modelos algorítmicos. De esta forma, podemos definir algunos conceptos que habitualmente se trasladan a la doctrina jurídica, como el *human in the loop*. No obstante, este acercamiento tiene una utilidad limitada desde un punto de vista normativo[218]. Por ello, se realiza posteriormente un examen de la cuestión desde el ámbito normativo, en el que se pone de manifiesto que es habitual encontrar la intervención humana como mecanismo de gobernanza de los sistemas automatizados -aunque adolecemos de una sistematización adecuada de las distintas formas de intervención posibles-. Por último, es absolutamente relevante cómo las instituciones de la UE, a la hora de abordar la regulación de los sistemas de IA de alto riesgo, han adoptado como requisito indispensable la "supervisión humana" de estos sistemas y han ido definiendo este concepto y su eventual aplicación normativa.

2.1. Acercamiento técnico a la intervención humana: un análisis limitado.

La literatura jurídica sobre la intervención humana en los sistemas automatizados ha estado y está muy influenciada por el acercamiento técnico a la cuestión. Por ello es

[217] Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales, «Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (CCW/GGE.1/2018/3)». Disponible en: https://undocs.org/en/CCW/GGE.1/2018/3

[218] Por ejemplo, en el tercer capítulo se expone cómo las figuras human in/out of the loop pueden ser útiles a la hora de discernir en qué punto el ordenamiento jurídico obliga a la intervención humana y, por ende, a discernir cómo los resultados algorítmicos despliegan sus efectos. Sin embargo, dicha clasificación no aporta nada a efectos de comprender el significado normativo de la propia intervención. Vid. Apartado 1. Derecho a la intervención humana. Capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica.

habitual encontrar referencias a figuras como: *human in the loop, human on the loop, human in command o human out of the loop*.

En este sentido, en primer lugar -también porque es la figura más utilizada en el ámbito normativo-, es relevante determinar qué significa mantener a un *human in the loop* en el sistema. Cabe señalar que hay dos concepciones principales de lo que es realmente el *human in the loop*.

Según una concepción técnica, podría definirse como: «*el proceso por el cual siendo la máquina o el sistema informático incapaz de resolver un problema, requiere la intervención humana en las etapas de entrenamiento y prueba del desarrollo de un algoritmo, para crear un bucle de retroalimentación continuo que permita al algoritmo dar cada vez mejores resultados*»[219]. Esta concepción considera la interacción humano-máquina directamente en el propio sistema algorítmico.

En cambio, la concepción normativa de esta interacción tiene su origen en el enfoque centrado en el ser humano desarrollado por Sheridan, el cual se centra en la intervención humana para mantenerla en el sistema decisorio, es decir, se trata de un concepto más amplio que trasciende de la visión del algoritmo como un mero sistema con datos de entrada y de salida, e incluye por lo general mantener al operador humano como la autoridad final sobre el sistema automatizado[220]. De esta manera, se asegura que el modelo en su conjunto involucra también la toma de decisiones humana y, por ende, los sistemas automatizados no son el único razonamiento tras el proceso de toma de decisiones[221]. Esta intervención puede tener un carácter más procedimental en ocasiones: por ejemplo, hacer avanzar o retroceder un caso determinado en el orden de preferencia establecido, o decidir qué casos merecen más recursos institucionales frente a otros. O un

[219] Vid. Singh Visen, «What is Human in the Loop Machine Learning: Why & How Used in AI?». Disponible en: https://medium.com/vsinghbisen/what-is-human-in-the-loop-machine-learning-why-how-used-in-ai-60c7b44eb2c0

[220] Vid. Sheridan, «Human centered automation: oxymoron or common sense?» Hay incluso definiciones más amplias de HITL que prácticamente incluyen figuras análogas como el HOTL: *sistemas que funcionen automáticamente en la mayoría de los casos, pero que prevean la posibilidad de un control humano en caso de error evidente*. En esta línea, Cranor, «A Framework for Reasoning about the Human in the Loop», 1-15.

[221] Wagner, «Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems», 108.

carácter dispositivo, pudiendo dar forma al resultado en unos términos concretos, como aceptar o denegar un crédito, o en términos de cómo se justifica dicho resultado[222].

Otra figura con ciertas similitudes con la anterior, dado que engloba también un amplio rango de sistemas que incorporan al ser humano en el ciclo de toma de decisiones manteniendo un control más o menos intenso sobre el proceso, es el *human on the loop*. Según el Grupo de expertos de alto nivel sobre inteligencia artificial (en adelante, AI-HLEG), *human on the loop* se refiere a la capacidad de que intervengan seres humanos durante el ciclo de diseño del sistema y en el seguimiento de su funcionamiento[223]. Bajo esta figura, la función del ser humano se limita a la vigilancia del funcionamiento del sistema; se asemeja más a una supervisión en tiempo real[224].

Por último, al margen de los sistemas híbridos cabe la posibilidad de que el ser humano se sitúe fuera del sistema decisorio. Ello puede significar que estamos ante un sistema plenamente automatizado, pero también es posible que se incluyan figuras que actúan al margen o de forma posterior al sistema algorítmico. Por un lado, *human in command* es la capacidad de supervisar la actividad global del sistema de IA -*incluyendo, desde un punto de vista más amplio, sus efectos económicos, sociales, jurídicos y éticos*- así como la capacidad de decidir cómo y cuándo utilizar el sistema en una situación determinada[225]. Lo cual, *puede incluir la decisión de no utilizar un sistema de IA en una situación particular, establecer niveles de discrecionalidad humana durante el uso del sistema o garantizar la posibilidad de ignorar una decisión adoptada por un sistema*[226]. Por otro lado, con *human out of the loop*, nos podemos referir a casos en los cuáles el funcionamiento del sistema es revisado a posteriori, es decir, el sistema produce efectos de forma automatizada e inmediata, sin perjuicio de que después, exista la posibilidad de revisar o modificar a través de un ser humano las decisiones adoptadas.

---

[222] Brennan-Marquez, Levy, y Susser, «Strange Loops: Apparent Versus Actual Human Involvement in Automated Decision Making», 749.

[223] Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (HLEG-AI), *Directrices éticas para una IA fiable*, 22.

[224] Fischer et al., «In-the-loop or on-the-loop? Interactional arrangements to support team coordination with a planning agent», 1.

[225] Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (HLEG-AI), *Directrices éticas para una IA fiable*, 22.

[226] Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (HLEG-AI), 22.

A pesar de que estas clasificaciones pueden resultar útiles y las encontramos habitualmente en la doctrina, tratar de realizar una sistematización conceptual de la intervención humana a partir de estas figuras tiene un escaso valor normativo[227] y, en el mejor de los casos, únicamente produce confusión[228]. Ciertamente, una conceptualización más satisfactoria ha de poder responder a preguntas relevantes acerca de la interacción humana respecto del sistema automatizado. Desde esta perspectiva analizada, puede decirse quién es la persona que ostenta el control del sistema y cuáles son los momentos o herramientas para la intervención humana, no obstante, hay otras preguntas que han de ser abordadas acerca cuál sería el nivel adecuado de control y qué mecanismos o medidas son necesarias para garantizar dicho control[229].

2.2. Intervención humana en el contexto regulatorio

En el contexto regulatorio es habitual encontrar un enfoque basado en el principio de precaución y que se ha traducido habitualmente en el contexto normativo europeo en la utilización del elemento humano con carácter preceptivo como mecanismo de gobernanza para abordar los efectos -no deseados- de la automatización. De hecho, se considera que los seres humanos son cruciales para evitar correlaciones inadecuadas y, por lo tanto, para garantizar la equidad en la extracción de datos[230]; pero no solo eso, la intervención humana también ha sido justificada para evitar que los seres humanos sean tratados de

[227] En esta línea, Enarsson, Enqvist, y Naarttijärvi, «Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts», 126..

[228] Así lo expresa Jiménez Segovia en relación al debate en el Derecho Internacional Humanitario sobre los : *La desorientación y confusiones causadas por la perspectiva de sistematización conceptual basada en los sistemas de armas semiautónomos, con autonomía supervisada y completamente autónomas (in, on, out the loop, respectivamente) originó que pronto surgieran voces alternativas que proponían salir del bucle control humano/autonomía, poniendo el foco de atención, no en la autonomía de los sistemas en general desde un punto de vista estrictamente técnico, ni tampoco en los niveles de complejidad de la máquina (automático, automatizado y autónomo), sino en la autonomía de las armas para ejecutar las específicas y verdaderamente relevantes funciones (relacionadas con el uso de la fuerza), a efectos de evaluar su aptitud para dar cumplimiento a los principios de derecho internacional humanitario y que conectan a estas armas con las cuestiones éticas más importantes*. Jiménez-Segovia, «Los sistemas de armas autónomos en la Convención sobre ciertas armas convencionales: Sombras legales y eticas de una autonomía ¿bajo el control humano?», 9-10.

[229] Vid., por ejemplo, las preguntas de la versión piloto de la Lista de Evaluación para una IA Fiable de las Directrices Éticas preparadas por el AI-HLEG. Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (HLEG-AI), *Directrices éticas para una IA fiable*, 35.

[230] Favaretto, De Clercq, y Elger, «Big Data and discrimination: perils, promises and solutions. A systematic review», 21.

manera totalmente automatizada, proporcionando así la restauración de la dignidad humana[231].

Es decir, el establecimiento de distintas clases de intervención humana como mecanismo de gobernanza estaría en principio motivado, no solo para evitar perjuicios en distintos bienes jurídicos, sino también porque, las decisiones totalmente automatizadas son consideradas *per se* perjudiciales para los derechos humanos, en particular para la dignidad y la autodeterminación humanas[232]. Dicho de otro modo, la intervención está principalmente motivada, o bien porque se cree que aporta una mejora en la toma de decisiones, o bien porque se cree que aporta otros valores a la toma de decisiones dignos de protección jurídica, como la licitud o la dignidad[233].

Ahora bien, no es habitual encontrar en la regulación una mención específica a los fines a los que obedece la introducción de la intervención humana en cualquiera de sus formas, y tampoco la doctrina ha sistematizado los distintos mecanismos de gobernanza que podemos encontrar en la regulación vigente. Por ello, incluyo aquí la investigación realizada por Green y que da pasos importantes en esta tarea. En su análisis, que incluye el análisis de normativa, documentos o declaraciones de carácter político-jurídico y resoluciones judiciales, Green identifica tres aproximaciones distintas en el ámbito normativo[234]:

- Primera aproximación. Restricción de las decisiones exclusivamente automatizadas: este acercamiento normativo estaría presente en el artículo 22(1) del RGPD, pero Green encuentra además varios ejemplos en Derecho comparado -Argentina, Mauritania, Kenia, Brasil, Sudáfrica o los EEUU, entre otras- que también recogen una prohibición o restricción directa de las decisiones que se toman por medios "exclusivamente" automatizados. En todos estos casos observa unos derechos o garantías adicionales para los casos en los que se levanta la restricción, encontrando en la mayoría el derecho de los sujetos de las decisiones

[231] Jones, «The right to a human in the loop: Political constructions of computer automation and personhood», 30.

[232] Ambos fundamentos resultan discutibles, aunque por el momento no se abordarán de forma crítica.

[233] Brennan-Marquez, Levy, y Susser, «Strange Loops: Apparent Versus Actual Human Involvement in Automated Decision Making», 746.

[234] Vid. Green, «The Flaws of Policies Requiring Human Oversight of Government Algorithms».

exclusivamente automatizadas a obtener una intervención humana a posteriori. En otras palabras, después de que alguien haya sido objeto de una decisión exclusivamente automatizada, puede solicitar que un humano inspeccione y considere la posibilidad de modificar esa decisión[235].

- Segunda aproximación. Énfasis en la discrecionalidad humana: en este caso, subrayan que las decisiones deben incorporar discrecionalidad humana, lo que proporciona protección contra los peligros potenciales de las decisiones automatizadas identificados habitualmente y relacionados con la vulneración de derechos fundamentales. Es decir, este acercamiento normativo justifica la introducción de la discrecionalidad humana -y su superioridad frente al juicio algorítmico- para la protección de los derechos fundamentales. Se trataría de un acercamiento con una presencia destacable en la aplicación de sistemas algorítmicos en el ámbito penal y procesal penal[236].

- Tercera aproximación. Requerimiento de una intervención humana "significativa": en este último acercamiento hay un requerimiento cualitativo de la intervención humana y puede encontrarse, entre otros, en varios documentos de instituciones de la UE a las que haré referencia en el próximo apartado. Aunque en ninguno de los documentos analizados encuentra una descripción detallada de la "cualificación" exigida, Green identifica tres componentes centrales: (1) la capacidad de los decisores humanos de no seguir el criterio algorítmico, imponiendo el suyo propio por competencia y autoridad dadas; (2) la capacidad de los decisores humanos para comprender cómo el sistema opera y adopta decisiones; (3) y la necesidad de que los decisores humanos no confíen ciegamente en el sistema, considerando toda la información relevante y necesaria para adoptar la mejor decisión posible[237].

Enlazando con este último acercamiento, un ámbito del Derecho que tiene un mayor desarrollo sectorial acerca de la regulación de la intervención humana en sistemas autónomos -y resulta, por ende, de interés en este punto-, es el Derecho Internacional

[235] Green, 9-10.

[236] Green, «Data Science as Political Action: Grounding Data Science in a Politics of Justice», 11-12.

[237] Green, 12-13.

Humanitario que ha venido desarrollando el concepto de Control Humano Significativo para el uso de los SAAL[238]. Adams acuñó el término en un artículo en el que planteaba cómo mantener el control humano sobre sistemas de armas autónomos que proliferarían con el paso del tiempo[239], tal y como certifican los sistemas utilizados en la actualidad gracias a los avances en nanotecnología, robótica e inteligencia artificial[240]. Es la propia definición de estos sistemas lo que requiere de este concepto de lo "significativo", dado que el elemento diferenciador de este tipo de armas respecto a otras es la eliminación, en parte o en su totalidad, de lo que la comunidad internacional ha denominado control humano significativo[241]. Tal y como de Sio y van den Hoven ponen de manifiesto, desde el inicio de este debate, la literatura ha venido reconociendo un problema respecto de este principio del control significativo que se traslada actualmente a otros sistemas automatizados: el problema es que carecemos de una teoría sobre el significado de este término[242].

La creciente preocupación por el uso de estos sistemas y la movilización de distintas organizaciones no-gubernamentales, de la ciencia y de la sociedad civil en general, ha propiciado que el debate sobre la regulación de los SAAL sea un aspecto central en el seno de Naciones Unidas y, en particular, desde el punto de vista del Derecho Internacional Humanitario (DIH). En esta línea son muy destacables los trabajos del Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales[243] en el marco de la Convención sobre Prohibiciones o Restricciones del Empleo de Ciertas Armas Convencionales que puedan

[238] En este sentido, el Grupo Europeo de Ética en la Ciencia y las Nuevas Tecnologías (en adelante EGE) ha destacado la importancia de este concepto como precedente a la hora de garantizar la supervisión humana para los sistemas autónomos. E igualmente se ha destacado por parte de la doctrina. Vid. European Group on Ethics in Science and New Technologies, «Future of Work, Future of Society»; Methnani et al., «Let Me Take Over: Variable Autonomy for Meaningful Human Control», 1; Romeo Casabona, «Criminal responsability of robots and autonomous artificial intelligent systems?»; Wagner, «Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems», 119.

[239] Adams, «Future Warfare and the Decline of Human Decisionmaking».

[240] Pérez Calvo, «Debate internacional en torno a los sistemas de armas autónomos letales. Consideraciones tecnológicas, jurídicas y éticas», 459.

[241] Pérez Calvo, 459.

[242] de Sio y van den Hoven, «Meaningful Human Control over Autonomous Systems: A Philosophical Account», 3.

[243] Este Grupo fue establecido en la Decisión 1 del documento final (CCW/CONF.V/10) de la Quinta Conferencia de Examen de las Altas Partes Contratantes en la Convención sobre Prohibiciones o Restricciones del Empleo de Ciertas Armas Convencionales que Puedan Considerarse Excesivamente Nocivas o de Efectos Indiscriminados, celebrada en Ginebra del 12 al 16 de diciembre de 2016.

considerarse excesivamente nocivas o de efectos indiscriminados, hecha en Ginebra el 10 de octubre de 1980[244].

2.2.1. La supervisión humana en las propuestas europeas de regulación de los sistemas de IA de alto riesgo

Siguiendo con lo desarrollado, existen diferentes enfoques normativos para abordar el contexto tecnológico-disruptivo definido en esta investigación. El enfoque regulador europeo es más cauteloso que en otras partes del mundo, lo que significa que la ausencia de certezas sobre el daño que pudiere producir una tecnología, no se considera generalmente como una razón para que el despliegue de la misma siga adelante; en cambio, los posibles beneficios y daños deben explorarse *ex ante* de la manera más profunda y exhaustiva posible[245].

Como muestra de ello, resulta muy interesante ver el acercamiento que han adoptado sobre esta cuestión las instituciones de la UE a la hora de abordar la regulación de los sistemas de IA.

En febrero de 2020, la Comisión Europea publicó el Libro Blanco sobre la IA. A pesar de que este documento no es jurídicamente vinculante, basándose en el mismo la Comisión se propone construir un sólido marco normativo europeo para el desarrollo una IA fiable (lícita-ética-robusta), con lo que resulta pertinente examinar cómo entiende la intervención humana en este contexto. En primer lugar, cabe señalar que el principal rol humano concebido en este documento es la "supervisión humana". Este concepto está

[244] Esta Convención negociada en el seno de Naciones Unidas es un destacado instrumento de DIH, y ante el mencionado debate acerca de las tecnologías emergentes en la esfera de los SAAL, el Grupo de Expertos antes mencionado consideró en el Informe del periodo de sesiones de 2018 que la Convención ofrecía un marco adecuado para abordar esta cuestión. Así se expresaba en el Apartado 30. Vid. Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales, «Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (CCW/GGE.1/2018/3)». Los trabajos de este grupo pusieron de manifiesto, una vez más, la necesidad de (a) desarrollar el concepto y aplicación del elemento humano; (b) definir el tipo y grado de intervención para favorecer el cumplimiento normativo; y (c) establecer la intervención humana más allá de las fases de implementación del ciclo de vida del sistema automatizado. Un análisis más amplio sobre el Control Humano Significativo en el DIH, y en particular sobre los resultados del Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales, puede encontrarse en Obregón Fernández y Lazcoz Moratinos, «La supervisión humana de los sistemas de inteligencia artificial de alto riesgo. Aportaciones desde el Derecho Internacional Humanitario y el Derecho de la Unión Europea».

[245] Recht et al., «Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations».

tomado tanto del enfoque centrado en el ser humano desarrollado por la Comunicación de la Comisión sobre generar confianza en la inteligencia artificial centrada en el ser humano[246], como de los siete requisitos para una IA fiable proclamados en las Directrices Éticas preparadas por el Grupo Independiente de Expertos de Alto Nivel sobre IA (HLEG-AI, en adelante)[247]. Su objetivo es contribuir a garantizar que un sistema de IA no socave la autonomía humana ni cause otros efectos adversos –refleja el doble fundamento ya mencionado–, y el Libro Blanco lo incluye entre los requisitos legales obligatorios para las aplicaciones de IA de alto riesgo que se impondrán a los agentes pertinentes en el futuro marco normativo.

En cuanto al contenido de la supervisión humana como requisito, la Comisión afirma que sólo se puede lograr una IA fiable mediante una participación adecuada de los seres humanos en relación con las aplicaciones de IA de alto riesgo[248]. Puede requerirse un tipo y grado diferente de participación humana dependiendo del uso previsto de la IA y sus potenciales efectos. Fuera del alcance de esta clase de intervención quedarían las manifestaciones indirectas de intervención humana[249]. Por consiguiente, la Comisión entiende aquí que el requisito de la supervisión humana se logra mediante mecanismos de gobernanza que requieren diferentes tipos de participación o intervención humana en el proceso de adopción de decisiones, y ofrece una lista no exhaustiva de esos mecanismos de gobernanza[250]:

- *El resultado del sistema de IA no es efectivo hasta que un humano no lo haya revisado y validado (por ejemplo, la decisión de denegar una solicitud de prestaciones de seguridad social solo podrá adoptarla un ser humano)*. Este supuesto define el *human in the loop*. Anteriormente, el HLEG-AI y la propia

[246] Vid. Comisión Europea, Comunicación de la Comisión «Generar confianza en la inteligencia artificial centrada en el ser humano». Bruselas, 8.4.2019. COM(2019) 168 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=CELEX:52019DC0168

[247] Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (HLEG-AI), *Directrices éticas para una IA fiable*.

[248] Comisión Europea, Libro Blanco sobre la inteligencia artificial, 21.

[249] Tales como los procedimientos de prueba, inspección o certificación necesarios para verificar el cumplimiento de los requisitos obligatorios. En un sentido amplio podrían considerarse formas de intervención humana, pero estos mecanismos de gobernanza no están integrados activamente en el proceso de adopción de decisiones.

[250] Para mí, esta lista ofrecida por la Comisión no resultaba satisfactoria, ya que incidía nuevamente sobre el acercamiento técnico tradicional a la cuestión y no consiguió aportar un concepto sólido de la supervisión humana como requisito normativo obligatorio.

Comisión ya consideraban el *human in the loop* como un mecanismo de gobernanza para lograr la supervisión humana y ha sido recogido por la normativa europea, por ejemplo y entre otros, en el RGPD mediante la prohibición de la toma de decisiones automatizada en su artículo 22.

- *El resultado del sistema de IA es inmediatamente efectivo, pero se garantiza la intervención humana posterior (por ejemplo, la decisión de denegar una solicitud de tarjeta de crédito puede tramitarse a través de un sistema de IA, pero debe posibilitarse un examen humano posterior)*. Es posible imponer una intervención *out of the loop*, esto es, desde un punto de vista normativo se reconoce la legitimidad directa de un resultado algorítmico y sus efectos, garantizando a su vez una intervención humana que pueda revertir dicho resultado a posteriori. Este es el caso de la salvaguarda contenida para la toma de decisiones automatizada legítima en el párrafo 3 del artículo 22 RGPD, que se examinará también en el próximo capítulo.

- *Se realiza un seguimiento del sistema de IA mientras funciona y es posible intervenir en tiempo real y desactivarlo (por ejemplo, un vehículo sin conductor cuenta con un procedimiento o botón de apagado para las situaciones en las que un humano determine que el funcionamiento del vehículo no es seguro)*. Este ejemplo recoge la figura del *human on the loop*, para ello, recoge también un procedimiento o botón de apagado, es decir, una característica técnica. Aquí puede observarse con más claridad que en los ejemplos anteriores las diferencias entre los acercamientos técnicos (*human in/on/out of the loop*) y el acercamiento normativo. Desde un punto de vista normativo, requerir que un sistema posibilite un seguimiento mientras funciona y sea posible intervenir en tiempo real y desactivarlo, implica que desde el diseño y desarrollo se imponga a su vez la necesidad de incorporar las características técnicas necesarias con dicho fin.

- *En la fase de diseño, se imponen restricciones operativas al sistema de IA (por ejemplo, un vehículo sin conductor dejará de funcionar en determinadas condiciones de visibilidad reducida en las que los sensores sean menos fiables, o mantendrá una cierta distancia con el vehículo que lo preceda en una situación dada)*. En este último ejemplo se ilustra igualmente esa diferencia entre la esfera técnica y normativa. Que las restricciones operativas devuelvan el control al

agente humano o directamente detengan el sistema automatizado, desde un punto de vista normativo, requiere que las restricciones sean impuestas desde el diseño y el desarrollo de los sistemas por medio de características técnicas.

Aunque con menor desarrollo, 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[251] también recalca la importancia de la supervisión humana como mecanismo de gobernanza, estableciendo como un principio ético de obligado cumplimiento el desarrollo de una IA antropocéntrica, antropogénica y controlada por seres humanos en su artículo séptimo, cuyo contenido puede resumirse en la garantía de una supervisión humana integral (apartado primero) que, en todo caso, permita el restablecimiento del control humano cuando sea necesario (apartado segundo).

Dicho apartado primero establece que debe de garantizarse en todo momento, haciendo referencia a las fases de desarrollo, despliegue y uso, una supervisión humana integral. No opta, en definitiva, por ningún mecanismo de intervención humana en particular, sino que se limita a regular que la supervisión humana debe darse en todo el ciclo de vida de la IA.

El Considerando 10 sí incluye una referencia algo más concreta, declarando que las decisiones adoptadas deben ser objeto de revisión, evaluación, intervención y control humanos significativos; es decir, menciona cuatro formas de participación humana (revisión, evaluación, intervención y control) y hace referencia al concepto "significativo" (*meaningful* en la literatura anglófona), que aparece también en la propia resolución del Parlamento en relación con los sistemas de armas autónomos letales (SAAL) -par. 89- y a la protección de la intimidad en la toma de decisiones por los poderes públicos -par. 69-. Sin embargo, este concepto "significativo" no aparece en el articulado[252].

En el apartado segundo, puede observarse que, aquí sí, el Parlamento Europeo considera necesario incluir un mecanismo de gobernanza específico que permita el restablecimiento

[251] 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)). Disponible en: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_ES.html

[252] Vid. 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)).

del control humano *cuando sea necesario, incluso mediante la alteración o la desactivación de dichas tecnologías*. Esta disposición tiene más cabida en aquellas tecnologías cuyo despliegue se produce en una dimensión física (como los SAAL o los vehículos autónomos) y parece que su inclusión es un tanto superflua respecto del primer apartado. A fin de cuentas, determinar cuándo es *necesario* restablecer el control humano en el despliegue de un sistema automatizado, parece indisolublemente ligado a asegurar una supervisión humana integral del sistema.

Más recientemente, en abril de 2021, la Comisión ha vuelto a enfatizar el peso de la supervisión humana para los sistemas de IA de alto riesgo en la propuesta de Reglamento AIA[253]. Y así, recoge en su artículo 14 la supervisión humana como requisito obligatorio de los sistemas de IA de alto riesgo[254], debiendo éstos ser diseñados y desarrollados de forma que se garantice que puedan ser supervisados eficazmente por agentes humanos durante el uso y despliegue de los mismos. El modelo regulatorio por el que ha optado la Comisión responsabiliza a los proveedores de los sistemas del cumplimiento de los requisitos obligatorios como la supervisión humana, poniendo definitivamente el enfoque normativo sobre las fases de diseño y desarrollo de estas tecnologías que habían permanecido fuera del análisis doctrinal jurídico y de las iniciativas políticas anteriores[255].

Resulta oportuno subrayar dos características fundamentales sobre las que se construye la supervisión humana como requisito obligatorio para los sistemas de IA de alto riesgo. Por un lado, el conjunto de obligaciones está dirigido al proveedor en la fase de diseño y desarrollo de los modelos, reforzando un modelo de supervisión durante todo el ciclo de vida de los sistemas. Por otro lado, no se opta por un mecanismo de gobernanza concreto, sino por cualificar dicha supervisión, requiriendo de medidas que puedan procurar una

[253] Comisión Europea, Propuesta de Reglamento por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial). Bruselas, 21.4.2021. COM(2021) 206 final. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex:52021PC0206

[254] La versión en lengua castellana de esta propuesta ha traducido "human oversight" como "vigilancia humana". A mi parecer hay motivos de peso para tomar como referencia supervisión humana y no vigilancia. En primer lugar, esta traducción se aparta de los precedentes mencionados, el Libro Blanco de la propia Comisión y la propuesta del Parlamento. En segundo lugar, en la misma propuesta también se traduce "market surveillance" como "vigilancia del mercado" –esta traducción sí parece ajustada–, provocando una reiteración del término vigilancia en contextos distintos.

[255] Lehr y Ohm, «Playing with the data: what legal scholars should learn about machine learning», 655 y ss.

supervisión *efectiva* por personas físicas[256], que consiste en que las personas a quienes se encomiende la supervisión del sistema deben ser capaces de, entre otros, entender por completo las capacidades y limitaciones del sistema, ser conscientes del sesgo de automatización, interpretar correctamente la información de salida del sistema, desestimar, invalidar o revertir dicha información o interrumpir el sistema accionando un botón específicamente destinado a tal fin (art. 14(4) AIA).

El desafío normativo actual consiste en determinar cuál es la participación humana adecuada para las numerosas aplicaciones de los modelos algorítmicos y sus distintos ámbitos de aplicación para la toma de decisiones automatizada (¿qué forma debe adoptar la participación?; ¿cuál es el fundamento de esta participación?; ¿para qué sistemas debe ser obligatoria?)[257]; y en cómo determinar en el ámbito normativo la cualificación que se exija para dicha participación (¿qué es adecuado, efectivo o significativo?).

Ahora bien, a la hora de afrontar este desafío nos encontramos con una dificultad jurídica importante antes mencionada, y es que desde el Derecho no se ha estudiado de forma sistematizada la intervención humana sobre los modelos automatizados de toma de decisiones -quizás el esfuerzo más relevante en este sentido es la inclusión de la supervisión humana como requisito obligatorio para los sistemas de IA de alto riesgo en las iniciativas regulatorias mencionadas[258]-. Ciertamente, la supervisión significativa de los modelos algorítmicos que trascienden nuestra capacidad cognitiva supone un reto formidable, especialmente para quienes no trabajan con un modelo algorítmico a diario[259].

[256] Una vez más, varía la cualificación, en el Libro Blanco se optaba por una participación humana adecuada y en la propuesta del Reglamento por una supervisión humana significativa.

[257] Para el AI-HLEG los mecanismos de supervisión deberán apoyar otras medidas de seguridad y control, y añade: (...) Si el resto de las circunstancias no cambian, cuanto menor sea el nivel de supervisión que pueda ejercer una persona sobre un sistema de IA, mayores y más exigentes serán las verificaciones y la gobernanza necesarias. Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (HLEG-AI), *Directrices éticas para una IA fiable*, 20.

[258] De algún modo, cuando se reclama en este contexto una supervisión humana efectiva/adecuada, parece asumirse que hasta ahora se ha mantenido una supervisión humana efectiva/adecuada sobre las tecnologías y modelos algorítmicos que utilizamos y utilizábamos para la toma de decisiones. No obstante, los efectos de la automatización no son exclusivos de los sistemas de IA más avanzados -por ejemplo, el sesgo de complacencia o automatización se describió hace décadas-. Es posible que el desarrollo y utilización de esta clase de tecnologías apremie a la hora de sistematizar y requerir el cumplimiento de una supervisión humana adecuada, lo cual no quiere decir que no fuera ya necesario anteriormente.

[259] Kolkman, «The (in)credibility of algorithmic models to non-experts», 106.

3. Sesgos en la toma de decisiones automatizada basada en la elaboración de perfiles

Las investigaciones de KAHNEMAN y TVERSKY acerca de los sesgos cognitivos comenzaron a desvelar las diferentes distorsiones y alteraciones que se producen en el cerebro humano al procesar información[260]. Una de las mayores promesas de los modelos algorítmicos[261] es la de reducir considerablemente los efectos nocivos de estos sesgos, contribuyendo, por ejemplo, a aportar consistencia a la toma de decisiones no dependiendo de factores como el cansancio o el hambre que afectan al juicio de los agentes humanos. Sin embargo, tal y como ha sido documentado en múltiples áreas[262], los algoritmos no están exentos de sesgos que pueden derivar en resultados discriminatorios, pudiendo reproducir e incluso amplificar los sesgos humanos.

En el ámbito normativo las referencias a los sesgos son más bien escasas y, en todo caso, muy recientes. Como reflejo de ello, el RGPD de 2016 en su Considerando 71 hace referencia a la corrección de inexactitudes en los datos y de reducción del riesgo de error al utilizar procedimientos matemáticos o estadísticos para la elaboración de perfiles que pudieran dar lugar a efectos discriminatorios, entre otros[263]. No obstante, en ningún momento hace referencia a los sesgos como un riesgo o factor a mitigar, como sí lo hace en 2021 la propuesta de Reglamento AIA de la Comisión Europea, como veremos. Ello pone de manifiesto que el interés normativo por los sesgos es reciente y está vinculado al igualmente reciente debate jurídico sobre la regulación de los modelos algorítmicos de

[260] Kahneman y Tversky, «Subjective probability: A judgment of representativeness».

[261] En realidad, tal y como apuntan KLEINBERG et al., este ha sido el objetivo de todas las clases de predicciones basadas en la estadística, pero los algoritmos de aprendizaje automático han renovado el interés por la cuestión aportando un cambio de paradigma en la forma en la que se realizan dichas predicciones. Vid. Kleinberg et al., «Human Decisions and Machine Predictions».

[262] Castelluccia y Le Métayer, «Understanding algorithmic decision-making: Opportunities and challenges», 7.

[263] Considerando 71 RGPD: *(…) A fin de garantizar un tratamiento leal y transparente respecto del interesado, teniendo en cuenta las circunstancias y contexto específicos en los que se tratan los datos personales, el responsable del tratamiento debe utilizar procedimientos matemáticos o estadísticos adecuados para la elaboración de perfiles, aplicar medidas técnicas y organizativas apropiadas para garantizar, en particular, que se corrigen 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*.

aprendizaje automático actualmente utilizados por las tecnologías de inteligencia artificial.

A continuación, se expondrán los distintos orígenes y los distintos planos en los que pueden producirse y operar estos sesgos, tratando de aportar una definición y conceptualización consistente de este término desde un punto de vista normativo. En la clasificación inicial trataré de realizar un acercamiento técnico a la cuestión para comprender las principales categorías de sesgos y cómo se reproducen en los sistemas algorítmicos. A partir de aquí abordaré la dimensión ética desde la que se hace referencia habitualmente a los sesgos y su carácter "problemático" y potencialmente discriminatorio. Y, por último, es necesario entender qué dimensión jurídica adquieren los sesgos, con especial incidencia en la discriminación, y cómo pueden afrontarse los riesgos que presentan desde el ámbito normativo.

Esta conceptualización se hace necesaria a la luz de la definición aportada por el Parlamento Europeo en 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[264]. Como se verá, la definición de sesgo como un prejuicio o percepción personal resulta considerablemente limitada en este contexto. Es más, tratar de aportar una definición bajo la que amparar las distintas clases de sesgos y las dimensiones desde las que pueden analizarse es una tarea compleja y probablemente inútil.

Sí parece necesario precisar que los sesgos son un problema social, más que (o antes que) un problema tecnológico que puede solucionarse por medios tecnológicos[265]. A nivel técnico es evidente que los modelos algorítmicos de aprendizaje automático no pueden transformar 'malos' datos de entrada en 'buenos' datos de salida, precisamente por la forma en la que opera el aprendizaje automático. Estos modelos responden así a la conocida expresión *garbage in, garbage out* (o GIGO). Sin embargo, otras problemáticas son menos evidentes, como la producción de nuevos sesgos que no resultan directamente aprehensibles por el juicio humano; la producción de sesgos a partir de datos intachables

[264] Artículo 4 l): *«sesgo», toda percepción personal o social prejuiciosa de una persona o de un grupo de personas sobre la base de sus características personales*. Vid. 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)).

[265] Pot, Kieusseyan, y Prainsack, «Not all biases are bad: equitable and inequitable biases in machine learning and radiology», 2.

desde el punto de vista técnico; o incluso la producción de cambios en las conductas de los sujetos objeto de las predicciones de estos modelos [266].

3.1. Clases de sesgos

Realizar una clasificación exhaustiva de los sesgos representa un reto agravado por la ausencia de un consenso sobre una sistematización aceptada y una ordenación inequívoca de las categorías de sesgos[267]. Además, no es una tarea jurídica. Con lo cual, en este apartado me limitaré a recoger algunas encomiables tareas de sistematización que puedan permitirnos conocer de forma superficial la variedad de sesgos que pueden identificarse en el diseño, desarrollo e implementación de un modelo algorítmico para la toma de decisiones. Algunas de las categorías definidas a continuación son compatibles entre sí, unas diferencian los sesgos en función de su origen o del momento en que se producen, pero en general podemos considerar que la caracterización de los distintos tipos de sesgos es un elemento clave del objetivo analítico común para reconocer y remediar dichos sesgos en los sistemas algorítmicos existentes[268].

Friedman y Nissenbaum fueron pioneras en la consideración de los sesgos en los sistemas informáticos, preocupados por la forma en que estos sistemas discriminan de forma sistemática a ciertos individuos o grupos frente a otros por motivos que no son razonables o apropiados[269]. En su taxonomía diferenciaron entre tres tipos de sesgos que siguen considerándose de relevancia en la actualidad[270].

Los sesgos preexistentes ("*preexisting bias*") tienen su origen en las instituciones, prácticas y actitudes sociales preexistentes a la construcción del modelo, pueden tener su origen en sesgos individuales de las personas que desarrollan el modelo o ser producto de sesgos colectivos existentes en la sociedad y cultura en general, y se introducen en el sistema de forma implícita e inconsciente[271]. Los sesgos técnicos ("*technical bias*") surgen de limitaciones o aspectos técnicos que incluyen las limitaciones de las

[266] McQuillan, «Data Science as Machinic Neoplatonism», 258.

[267] Cirillo y Rementeria, «Bias and fairness in machine learning and artificial intelligence», 4.

[268] Simon, Wong, y Rieder, «Algorithmic bias and the Value Sensitive Design approach», 8.

[269] Friedman y Nissenbaum, «Bias in Computer Systems», 332.

[270] Vid. Simon, Wong, y Rieder, «Algorithmic bias and the Value Sensitive Design approach».

[271] Friedman y Nissenbaum, «Bias in Computer Systems», 333.

herramientas tecnológicas utilizadas -hardware y software-, el despliegue de sistemas algorítmicos desarrollados para un contexto distinto o de la formalización de constructos humanos como la cuantificación de elementos cualitativos[272]. Por último, los sesgos emergentes ("*emergent bias*") surgen posteriormente en el despliegue del sistema como resultado de nuevos conocimientos sociales o valores culturales cambiantes que no se incorporan o no pueden incorporarse al diseño del sistema, o de un desajuste entre los usuarios humanos -sus conocimientos y valores- concebidos en el diseño y desarrollo del sistema y la población real que lo utiliza[273].

La literatura se refiere habitualmente a "sesgos algorítmicos" como una categoría autónoma, entendiendo como tal la preocupación de que un sistema algorítmico no sea, en algún sentido, un mero transformador neutral de datos o extractor de información. No obstante, Danks y London consideran que esta noción de "sesgos algorítmicos" en el debate público ha contribuido a la confusión entre diferentes clases de sesgos y su impacto. Por ello, proponen una taxonomía sobre esta categoría que indague en la forma en que sesgos de distintos orígenes pueden reproducirse en los sistemas algorítmicos, permitiendo evaluar si un sesgo particular merece una respuesta y, en caso afirmativo, qué tipo de medidas correctoras deben aplicarse[274].

A mi modo de ver, una de las aportaciones más valiosas de este trabajo, sobre la que volveré en el próximo apartado, es la separación de la noción sesgo de una carga necesariamente negativa, puesto que dicha consideración depende del estándar sobre el que se evalúe la desviación que representa determinado sesgo[275]. De hecho, ello nos permite distinguir entre sesgos deseados -en tanto la desviación que introduce se valora como positiva- e indeseados -cuando es negativa-.

---

[272] Friedman y Nissenbaum, 334.

[273] Friedman y Nissenbaum, 335.

[274] Vid. Danks y London, «Algorithmic Bias in Autonomous Systems».

[275] Así lo explican: *The word 'bias' often has a negative connotation in the English language; bias is something to be avoided, or that is necessarily problematic. In contrast, we understand the term in an older, more neutral way: 'bias' simply refers to deviation from a standard. Thus, we can have statistical bias in which an estimate deviates from a statistical standard (e.g., the true population value); moral bias in which a judgment deviates from a moral norm; and similarly for regulatory or legal bias, social bias, psychological bias, and others. More generally, there are many types of bias depending on the type of standard being used*. Danks y London, 4692.

Aunque interrelacionados, la literatura distingue también entre sesgos estadísticos y sesgos cognitivos, y ambas categorías pueden a su vez traducirse en sesgos algorítmicos. Por sesgo estadístico entendemos la desviación sistemática del resultado de un procedimiento analítico con respecto a su valor real, excluyendo la contribución del azar[276]. Y dentro de esta categoría podemos distinguir a su vez distintos tipos de sesgos, como el sesgo de selección -producido cuando el grupo seleccionado para el análisis no es representativo de toda la población- o el sesgo de detección -que se produce cuando un evento es más probable de ser observado en un grupo específico de la población-[277].

Volviendo sobre los sesgos cognitivos definidos por primera vez por Kahneman y Tversky, pueden definirse como la desviación sistemática del procesamiento de información humano de algunos aspectos de la realidad objetiva, su aparición está estrechamente vinculada a la evolución de la cognición humana[278] y pueden ser también de muy distintos tipos[279]. Los sesgos algorítmicos están habitualmente causados, directa o indirectamente, por sesgos cognitivos que se ocultan y, a la vez, reproducen y refuerzan en sistemas aparentemente neutros; no obstante, si dichos sesgos cognitivos son apropiadamente identificados, los sistemas algorítmicos son concebidos como un eventual remedio para corregir las debilidades del juicio humano[280].

En sentido contrario, los sesgos cognitivos tienen, además, ramificaciones interesantes en un contexto de toma de decisiones automatizadas en el que las decisiones humanas se complementan -o incluso se sustituyen- por resultados de sistemas algorítmicos, es decir, en la implementación de estos sistemas[281]. La evidencia muestra que la automatización no se limita a suplir la actividad humana, sino que la modifica, a menudo de forma no

[276] Cirillo y Rementeria, «Bias and fairness in machine learning and artificial intelligence», 10.

[277] Vid. más en Cirillo y Rementeria, 11.

[278] Es muy interesante la conexión que establecen Cirillo y Rementería entre la generalización en el aprendizaje humano -como estrategia cognitiva que se apoya en la reducción de conceptos complejos a categorías simples- y la generalización que subyace en el aprendizaje automático de los sistemas algorítmicos, y cómo esas generalizaciones derivan en sesgos no deseados cuando responden a estereotipos o correlaciones burdas. Cirillo y Rementeria, 10.

[279] Al respecto vid. Benson, «Cognitive bias cheat sheet. An organized list of cognitive biases because thinking is hard». Disponible en: https://medium.com/better-humans/cognitive-bias-cheat-sheet-55a472476b18

[280] Simon, Wong, y Rieder, «Algorithmic bias and the Value Sensitive Design approach», 12.

[281] Simon, Wong, y Rieder, 12.

prevista por quienes desarrollan los sistemas[282]. Es el caso, entre otros, del sesgo de automatización[283] y del sesgo de complacencia[284].

En este sentido, parece evidente la utilidad de identificar no únicamente qué tipos de sesgos existen y coexisten, sino las fases en las que éstos se reproducen en los sistemas algorítmicos y producen su impacto. Un reciente estudio publicado por el Servicio de Estudios del Parlamento Europeo (EPRS) identifica de forma no exhaustiva -en palabras de sus autores esto sería inviable- cómo distintas clases de sesgos (cognitivos, culturales, estadísticos, de representación, etc.) pueden reproducirse en distintas fases de la toma de decisiones algorítmica, en el diseño y desarrollo del modelo (en las sub-fases arriba analizadas de definición del problema y establecimiento de sus objetivos; en la recolección de datos; entrenamiento del modelo; y validación) y en su implementación (vid. ilustración)[285].

[282] Parasuraman y Manzey, «Complacency and Bias in Human Use of Automation: An Attentional Integration», 381.

[283] El sesgo de automatización se caracteriza por la tendencia humana a confiar en exceso en máquinas supuestamente neutrales, siguiendo el criterio de las mismas sin buscar más información que las corrobore o contradiga, o incluso descartando la información de otras fuentes existentes, vid. Skitka, Mosier, y Burdick, «Does automation bias decision-making?». Este sesgo puede resultar benigno, incluso beneficioso, cuando un sistema algorítmico proporciona recomendaciones correctas (por ejemplo, acelerando la toma de decisiones), no obstante, da lugar a errores que pueden resultar incorregibles (por omisión o por comisión) cuando el sistema erra, sobre factores y causas que provocan el sesgo de automatización, vid. Parasuraman y Manzey, «Complacency and Bias in Human Use of Automation: An Attentional Integration», 391-97.

[284] Entendemos por sesgo de complacencia la creencia de los operadores humanos en la fiabilidad del sistema, lo que hace que no presten suficiente atención a la supervisión del proceso y a la verificación de los resultados del sistema, vid. Simon, Wong, y Rieder, «Algorithmic bias and the Value Sensitive Design approach», 13.. Esta complacencia se produce habitualmente en condiciones de acumulación de tareas, cuando las tareas manuales compiten con la tarea automatizada por la atención del operador humano, sobre más factores y causas que provocan esta complacencia, vid. Parasuraman y Manzey, «Complacency and Bias in Human Use of Automation: An Attentional Integration», 382-90.

[285] de Miguel Beriain et al., «Auditing the quality of datasets used in algorithmic decision-making systems».

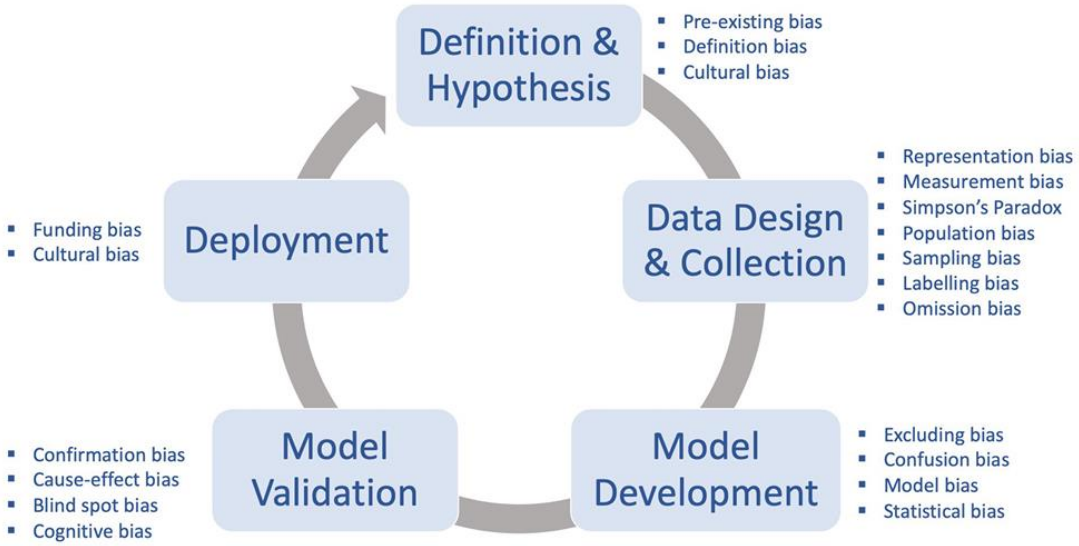

Ilustración 2. Fuente: De Miguel et al. 2022, p. 23.

3.2. Distintas dimensiones desde las que entender y analizar los sesgos en la toma de decisiones automatizada: más allá de la estadística

La multitud de sesgos que pueden reproducirse en los modelos algorítmicos teniendo repercusión en las inferencias y la toma de decisiones resultantes, pueden analizarse a su vez desde tres planos distintos.

En un primer lugar, encontramos un plano estadístico o matemático desde el que abordar los sesgos. En este plano los sesgos no tienen por qué representar un "mal" o un "bien" necesariamente, y se refiere a cualquier tipo de error o distorsión que se encuentra con el uso del análisis estadístico[286]. En términos estadísticos, un sesgo algorítmico puede ser entendido desde una concepción neutral, como una desviación de lo estándar cuyo valor ético o normativo no viene predeterminado[287].

Ahora bien, los sesgos no serían tan "populares" si no fuese por la repercusión que pueden tener desde un plano ético. Los sesgos (re)configuran la distribución de bienes, servicios, riesgos y oportunidades, o incluso el acceso a la información, de forma moralmente

[286] Piedmont, «Bias, Statistical BT - Encyclopedia of Quality of Life and Well-Being Research».

[287] Ello no quiere decir, por supuesto, que los fundamentos estadísticos sobre los que se examinan dichos sesgos no tengan una carga ética propiamente.

problemática[288]. Desde este punto de vista es posible que analicemos la producción de sesgos que desde el plano estadístico no se consideran tales. Es decir, puede que los sesgos se reproduzcan por un entrenamiento del modelo con datos incorrectos, irrelevantes o incompletos[289], en este caso el sesgo o desviación estadística puede reconfigurar la distribución de bienes, servicios, riesgos y oportunidades de forma moralmente problemática. No obstante, es posible que nos encontremos con una redistribución problemática sin necesidad de encontrar una desviación estadística, lo que se conoce como sesgos por una distribución desigual real de las variables[290], también conocidos como sesgos sociales. Aquí el problema viene perfectamente descrito por Hildebrandt: *the problem is not in the data but in the real world*[291].

Los datos con los que alimentamos los modelos proceden de nuestra sociedad, no siempre justa ni igualitaria, con lo que dichos datos, intachables desde un punto técnico/estadístico, pueden reflejar y reproducir situaciones desiguales y discriminatorias[292]. Además, el "aprendizaje" en los modelos de aprendizaje automático refuerzan la reproducción de esta clase de sesgos sociales, aprender asociaciones estadísticas de los datos de entrenamiento permite a la máquina producir respuestas correctas, pero a veces por razones equivocadas, dada la incapacidad de comprender realmente qué representan esos datos[293].

Estos dos primeros planos desde los que pueden entenderse los sesgos se corresponden, en realidad, con dos concepciones distintas de sesgos que habitualmente se utilizan de

[288] Hildebrandt, «The Issue of Bias. The Framing Powers of ML», 1.

[289] Entre este tipo de sesgos se incluyen también aquellos conjuntos de datos que pueden contar con registros individuales de excelente calidad en las que, sin embargo, un determinado colectivo está infrarrepresentado y, por ende, el algoritmo arroja resultados distorsionados respecto de dicho grupo. Barocas y Selbst, «Big Data's Disparate Impact», 684.

[290] Tomo esta denominación de Miró Llinares, «Inteligencia Artificial y Justicia: Más allá de los resultados lesivos causados por Robots»..

[291] Hildebrandt, «The Issue of Bias. The Framing Powers of ML», 4.

[292] Desde esta perspectiva, se señalan habitualmente las limitaciones a la hora de abordar la mitigación de sesgos de los sistemas algorítmicos, dado el vasto universo de cuestiones sociales, políticas, económicas y influencias estructurales en los sesgos algorítmicos que no pueden mitigarse sin cambios políticos que están fuera del alcance de soluciones reduccionistas basadas en la mitigación de sesgos sobre sistemas individuales. Vid. INNOCENCE PROYECT, «A proposal for identifying and managing bias within artificial intelligence -PUBLIC COMMENT ON DRAFT NIST Special Publication 1270-». Disponible en: https://www.nist.gov/artificial-intelligence/comments-received-proposal-identifying-and-managing-bias-artificial

[293] Mitchell, «Why AI is Harder than We Think», 3.

forma indistinta. Esta confusión terminológica tiene su origen en la lengua inglesa, puesto que *bias* se utiliza más habitualmente como prejuicio -por ende, con dicha carga moral problemática- que en su acepción estadística. En lengua castellana el diccionario de la RAE no recoge sesgo como prejuicio, no obstante, ello no ha evitado que la doctrina y las instituciones lo hayan utilizado con dicho significado en los últimos años al abordar la regulación de los modelos algorítmicos y las tecnologías de IA.

Ahora bien, la intersección entre ambas dimensiones también nos permite hablar de sesgos problemáticos y no problemáticos. Es más, introducir un sesgo en su sentido estadístico puede representar un bien desde una dimensión ética. Danks y London explican que algunos sesgos -desde su concepción como desviación estadística- pueden ser profundamente problemáticos, mientras que otros pueden ser componentes valiosos de un sistema fiable y éticamente deseable. Añaden que estas cuestiones no pueden resolverse de manera puramente tecnológica, ya que implican cuestiones cargadas de valores individuales y sociales como -utilizando el acceso al empleo como ejemplo- cuál debería ser la distribución de las oportunidades de empleo (independientemente de lo que realmente es en la actualidad), y qué factores deberían o no deberían influir en las perspectivas laborales de una persona[294].

La literatura ha señalado que la (re)configuración problemática -desde una perspectiva ética- de la distribución de bienes, servicios, riesgos y oportunidades está directamente relacionada con el objetivo mismo de estos modelos algorítmicos: discriminar[295]. Discriminar[296] para establecer diferentes precios en productos o primas de seguro. Discriminar para determinar el acceso a un empleo o a un tratamiento médico.

[294] Danks y London, «Algorithmic Bias in Autonomous Systems», 4692. Este punto de vista nos obliga a replantearnos qué tomamos por referencia cuando hablamos de una "ausencia de sesgos" y si ello tiene algún tipo de utilidad para la consecución de algoritmos "justos" o "precisos".

[295] Así lo expresan Veale y Binns: (…) *machine learning algorithms are supposed to discriminate between data points – that is why we use them – yet some logics of discrimination, even if predictively valid, are not societally acceptable*. Veale y Binns, «Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data», 2.

[296] Discriminar. Del lat. *discriminăre*. 1. tr. Seleccionar excluyendo. Versión electrónica de la 23ª edición del Diccionario de la lengua española (RAE)

Discriminar para ofrecer determinada publicidad o para aplicar beneficios penitenciarios. Lo cual no implica, necesariamente, que alguien haya sido discriminado[297].

Una vez más, la problemática se halla en el funcionamiento mismo de los algoritmos, puesto que la forma en la que producen resultados desiguales puede ser imperceptible y, desde luego, diferente de cómo los humanos producimos estos resultados. El potencial de estos modelos reside en establecer correlaciones y dibujar patrones entre múltiples datos que escapan a la cognición humana. Ello dificulta el control sobre los indicadores conocidos como *proxies* que, bajo una apariencia de neutralidad -el código postal es un ejemplo ampliamente utilizado-, correlacionan indirectamente con categorías problemáticas como la raza o el género. La utilización de grandes bases de datos facilita la interacción de estos *proxies* mientras que la opacidad epistémica de los algoritmos de aprendizaje automático dificulta el acceso a las correlaciones que sirven de base a los resultados algorítmicos[298].

Desde esta perspectiva ética, en el ámbito de las ciencias de la computación se ha trabajado los últimos años en el desarrollo del concepto de "*fairness*"[299] y en su aplicación para la toma de decisiones algorítmica, habitualmente traducido como "equidad". Para ello la comunidad de la ciencia de datos trabaja en distintas técnicas para la mitigación de sesgos que pueden estar presentes en distintas etapas del diseño, desarrollo e implementación de estos sistemas (*preprocessing-inprocessing-postprocessing*)[300].

No obstante, asumir que la equidad puede medirse y, por tanto, automatizarse resulta cuanto menos reduccionista; las propias discusiones en cuanto a cómo puede medirse esta

[297] Discriminar. Del lat. *discriminăre*. 2. tr. Dar trato desigual a una persona o colectividad por motivos raciales, religiosos, políticos, de sexo, de edad, de condición física o mental, etc. Versión electrónica de la 23ª edición del Diccionario de la lengua española (RAE)

[298] Heinrichs y Eickhoff, «Your evidence? Machine learning algorithms for medical diagnosis and prediction», 9.

[299] Este concepto ha tenido una influencia decisiva en la construcción de "algoritmos éticos", aunque otros como la transparencia o la responsabilidad (accountability) forman también parte de estos trabajos en la ciencia de datos. Uno de los foros más destacados en este contexto es la ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) que se celebra anualmente.

[300] de Laat, «Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?», 8-9. También, vid. Valdivia, Sánchez-Monedero, y Casillas, «How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness». Dichas etapas consisten fundamentalmente en: *The preprocessing techniques attempt to learn new representations of data to satisfy fairness definitions. The inprocessing methods involve modifying the classifier algorithm by adding a fairness criteria to the optimization problem. The postprocessing methods aim at removing discriminatory decisions after the model is trained*.

equidad reflejan que, de hecho, se trata de debates sobre un distinto entendimiento teórico de la equidad y de los valores que la conforman. Por ello, tratar de solidificar el concepto de equidad o *fairness* obviando su naturaleza contextual[301] y, particularmente, etiquetar modelos algorítmicos de tal modo resulta problemático[302].

3.2.1. Dimensión jurídica de los sesgos en la toma de decisiones automatizada

En último lugar, corresponde analizar el plano normativo. Adquiere aquí particular relevancia la discriminación, pero ello no quiere decir que la relevancia normativa de los sesgos tenga que limitarse al encaje de éstos en formas discriminatorias, los sesgos pueden llevar a resultados ilícitos sin necesidad de resultar discriminatorios. Es decir, los sesgos pueden perjudicar otros bienes jurídicos también protegidos por el ordenamiento jurídico.

Por ejemplo, pueden afectar al nivel de exactitud de los resultados algorítmicos sin necesidad de que se afecte a la forma en que se distribuyen bienes y servicios de forma igualitaria, no obstante, el descenso en el nivel de exactitud puede contribuir al aumento no justificado de daños en distintos bienes afectados por la utilización del sistema algorítmico. En este sentido, un sesgo que derive en resultados no-discriminatorios puede resultar tan perjudicial como un sesgo que resulte discriminatorio[303].

Por ello, desde un punto de vista normativo es relevante observar la clase de efectos perjudiciales que los sistemas algorítmicos pueden producir más allá de un posible trato desigual. El RGPD obliga al responsable del tratamiento en virtud del principio de exactitud a que los datos sean exactos y, si fuera necesario, actualizados, adoptando todas las medidas razonables para que se supriman o rectifiquen los datos personales que sean

[301] También sobre la naturaleza contextual de la discriminación en el plano normativo vid. Wachter, Mittelstadt, y Russell, «Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI».

[302] Así lo expresan Jacobs y Wallach: *Worse yet, by using such fairness definitions to label computational systems as "fair," we risk the adoption of these systems without any critical assessment because their fairness is assumed to be guaranteed*. Jacobs y Wallach, «Measurement and fairness», 384.

[303] A su vez, no todo sesgo que deriva en un resultado ilegal tiene que resultar ilegítimo desde un punto de vista ético, la normativa trata de estar alineada con determinados principios éticos, sin embargo, ello no implica que toda vulneración normativa sea éticamente problemática. Del mismo modo, como veremos a continuación con la generación de grupos *ad hoc*, puede que la legislación no proteja determinados casos que resultan problemáticos desde el plano ético.

inexactos con respecto a los fines para los que se tratan[304]. Con lo cual, la aparición y reproducción de sesgos en los modelos utilizados por los responsables del tratamiento pueden comprometer el cumplimiento de este principio. Asimismo, los sesgos son relevantes para distintas legislaciones sectoriales y, en particular, para las que exigen determinados estándares de calidad respecto de los productos dedicados al consumo. Es necesario evaluar cómo el funcionamiento de estos modelos y la reproducción de sesgos afecta al cumplimiento y control de estos estándares normativos.

Al margen de las cuestiones normativas aquí descritas, los sesgos en los sistemas automatizados han despertado un particular interés, justificado por las problemáticas éticas arriba señaladas, en relación al cumplimiento de la prohibición de discriminación.

3.2.1.1. Breve referencia a la normativa antidiscriminatoria

La normativa antidiscriminatoria, como conjunto de normas, jurisprudencia y doctrina que puede llegar a ser considerada una rama específica del ordenamiento jurídico[305], ha emergido en nuestro ordenamiento con un influjo determinante del Derecho comunitario[306]. Esta rama emerge a partir de la prohibición de discriminación contenida en el artículo 14 de la Constitución Española, que incorpora, a su vez, un listado abierto de causas de discriminación[307]: *«… sin que pueda prevalecer discriminación alguna por razón de nacimiento, raza, sexo, religión, opinión, o cualquier otra condición o circunstancia personal o social»*, y que limita no solo la actuación de los poderes públicos, sino también la autonomía de los sujetos privados[308].

[304] Artículo 5(1)(d) RGPD. Tal y como veremos en el tercer capítulo, este principio abarca todas las fases del proceso de elaboración de perfiles, inclusive al aplicarlo para tomar una decisión que afecta a una persona. Vid. Apartado 3. Derecho a impugnar las decisiones automatizadas. Capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica.

[305] Rey Martínez, «Igualdad y prohibición de discriminación: de 1978 a 2018», 129.

[306] Fernández López, «Artículo 14 CE: Igualdad ante la Ley y prohibición de discriminación», 6.

[307] Del mismo modo, ese carácter abierto es recogido por el apartado primero del artículo 21 de la Carta de los Derechos Fundamentales de la Unión Europea: *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*.

[308] Vid. Bilbao Ubillos, «Prohibición de discriminación y relaciones entre particulares».

En la normativa antidiscriminatoria es esencial la determinación de las llamadas "categorías sospechosas", esto es, características que se corresponden con elementos de la persona inmutables o que se sitúan en el núcleo mismo de la dignidad de la persona, por lo cual se atribuye un mayor desvalor a las decisiones que producen resultados desfavorables con base en dichos elementos[309], siendo habitual que dichas categorías se correspondan con grupos que han sufrido una situación histórico-cultural de desventaja que se perpetúa en la construcción de las estructuras sociales de poder actuales[310].

Con carácter general, el Derecho antidiscriminatorio distingue entre dos clases de discriminación.

La discriminación directa o de trato es aquella que se produce cuando una persona, por razón de su sexo, etnia, religión, etc., es tratada de manera menos favorable que otra en una situación comparable. En el contexto de los algoritmos de aprendizaje automático, la discriminación directa implica el establecimiento explícito de una categoría problemática como variable dentro del modelo, asignándole un valor menor respecto al resto; es, por así decirlo, la forma más burda de discriminación algorítmica, por lo que su concurrencia se considera excepcional[311], aunque no por ello absolutamente descartable[312].

Por el contrario, la discriminación indirecta o de impacto, por el contrario, se produce cuando una disposición, criterio o práctica aparentemente neutros sitúa a personas de una categoría "problemática" en desventaja particular con respecto a otras personas: «*salvo que dicha disposición, criterio o práctica pueda justificarse objetivamente con una finalidad legítima y salvo que los medios para la consecución de esta finalidad sean*

[309] Gerards, «The discrimination grounds of article 14 of the european convention on human rights», 114.

[310] Soriano Arnanz, «La propuesta de Reglamento de inteligencia artificial de la UE y los sistemas de alto riesgo», 5.

[311] Hacker, «Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law please cite to the final version forthcoming in: Common Market Law Review», 1152.

[312] En el conocido caso Wisconsin v. Loomis, la defensa alegó – sin éxito – que el uso del modelo algorítmico COMPAS constituía una discriminación por razón de sexo porque sus evaluaciones tenían en cuenta dicha categoría para determinar el riesgo de reincidencia de un individuo. Vid. Romeo Casabona, «Riesgo, procedimientos actuariales basados en inteligencia artificial y medidas de seguridad». Recientemente, Adams et al. han argumentado de forma sólida cómo los sistemas algorítmicos pueden incurrir en discriminación directa, rebatiendo a la doctrina mayoritaria y reclamando mayor atención sobre esta forma de discriminación. Vid. Adams-Prassl, Binns, y Kelly-Lyth, «Directly Discriminatory Algorithms».

adecuados y necesarios»[313]. Obedece, por tanto, a tres elementos: neutralidad, impacto sobre el colectivo y ausencia de justificación.

Esta clase de discriminación sí resulta de particular relevancia para los algoritmos de aprendizaje automático[314], parece encajar mejor con dichas nuevas formas de discriminación basadas en perfiles algorítmicos que gozan de indiscutible rigor matemático[315], en tanto que la neutralidad matemática de los códigos algorítmicos parece operar del mismo modo que la neutralidad dispositiva que genera resultados discriminatorios. Es decir, la discriminación indirecta permite poner el foco sobre los efectos de los sistemas algorítmicos, como un todo, más que sobre sus parámetros y reglas de funcionamiento internas[316].

Ahora bien, los retos para el derecho antidiscriminatorio son múltiples en este contexto. Primero, se han expuesto anteriormente las dificultades para el control de los llamados *proxies* que, sin pertenecer a las categorías problemáticas, pueden resultar igualmente discriminatorios. Esto es, la simple exclusión de la categoría problemática no es suficiente para asegurar que el modelo no utilizará otros indicadores alternativos que llevarán a la consideración correlativa de la categoría problemática[317]. Excluir, además de las categorías problemáticas, esos indicadores o *proxies* tampoco parece recomendable, puesto que se perdería mucha información útil disminuyendo sustancialmente la utilidad de los modelos[318]. Es decir, la eliminación de categorías problemáticas o de indicadores alternativos que puedan correlacionarse con las mismas, tiene una contrapartida en la precisión final del modelo.

Segundo, la doctrina considera que la opacidad algorítmica, en sus múltiples formas[319], supone una importante limitación para la aplicación efectiva de la normativa

[313] Art. 2.2 b) de la Directiva 2000/43/CE del Consejo, de 29 de junio de 2000, relativa a la aplicación del principio de igualdad de trato de las personas independientemente de su origen racial o étnico

[314] Hildebrandt, «Primitives of Legal Protection in the Era of Data-Driven Platforms», 272.

[315] Morente Parra, «Big Data o el arte de analizar datos masivos. Una reflexión crítica desde los derechos fundamentales», 252.

[316] Xenidis y Senden, «EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination», 21.

[317] Wellner y Rothman, «Feminist AI: Can We Expect Our AI Systems to Become Feminist?», 10.

[318] Hajian y Domingo-Ferrer, «Direct and Indirect Discrimination Prevention Methods», 243.

[319] Vid. Apartado 4. Opacidad en la toma de decisiones automatizada basada en la elaboración de perfiles, en este mismo capítulo.

antidiscriminatoria vigente[320] y a su vez una causa de intensificación de la discriminación[321]: la opacidad puede considerarse un problema en sí misma, pero la opacidad también dificulta descubrir la discriminación[322]. Si una persona al solicitar un crédito no conoce que ha sido objeto de una evaluación algorítmica, o si conociendo que ha sido objeto de dicha evaluación no puede acceder al razonamiento algorítmico por motivos legales -propiedad intelectual, secretos industriales etc.- o por su incapacidad para comprender el mismo, o no puede acceder a términos comparativos, esto es, evaluaciones de terceras personas por motivos de protección de datos, la aplicación de la normativa antidiscriminatoria al caso concreto será muy limitada.

Además, el acercamiento normativo actual del Derecho antidiscriminatorio solo aporta protección para aquellos modelos que discriminan sobre de las categorías problemáticas ya establecidas (sexo, orientación sexual, etnia, etc.), sin embargo, los algoritmos de aprendizaje automático están diseñados para crear nuevas categorías a partir de las inferencias que realizan, lo cual excede de las cuestiones relacionadas con la discriminación indirecta[323]. Estos grupos *ad hoc*[324], que son el resultado de patrones correlacionados por algoritmos de aprendizaje automático, podrían estar compuestos, por ejemplo, por personas que padecen asma ante un algoritmo que pretende predecir si pacientes con neumonía han de ser hospitalizados[325]. Dicho grupo no estaría actualmente cubierto por la normativa antidiscriminatoria, aunque pueden sufrir otros efectos perjudiciales[326].

El análisis inferencial de los algoritmos de aprendizaje automático amplía la gama de víctimas de actos discriminatorios, víctimas que podrían no corresponderse con las categorías actualmente previstas por la normativa y, por ende, el Derecho

[320] Soriano Arnanz, «La propuesta de Reglamento de inteligencia artificial de la UE y los sistemas de alto riesgo», 30-31.

[321] Heinrichs y Eickhoff, «Your evidence? Machine learning algorithms for medical diagnosis and prediction», 8.

[322] Zuiderveen Borgesius, «Strengthening legal protection against discrimination by algorithms and artificial intelligence», 1583.

[323] Mann y Matzner, «Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination», 4.

[324] Vid. sobre este concepto Mittelstadt, «From Individual to Group Privacy in Big Data Analytics».

[325] Ejemplo inspirado en la investigación realizada por Caruana et al., «Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission».

[326] Büchi et al., «The chilling effects of algorithmic profiling: Mapping the issues», 12.

antidiscriminatorio puede no proteger adecuadamente contra estos riesgos[327]. En definitiva, la evidencia sobre los sesgos en la toma de decisiones automatizada parece obligarnos a revisar algunos de los pilares fundamentales del aparato conceptual antidiscriminatorio, y también de la interpretación que de este aparato realizan los tribunales[328].

Como reflexión final a este apartado, conviene destacar que la discriminación en el uso de sistemas de toma de decisiones algorítmicas no se limita a la aparición de sesgos. Asumiendo que los sesgos pudiesen ser mitigados e incluso eliminados de los sistemas algorítmicos debido a altos estándares de calidad, validación en el entorno real y cumplimiento normativo. Un algoritmo libre de sesgos no quiere decir que su uso sea legítimo o que no incurra en discriminación alguna.

Los problemas de la automatización no terminan en los sesgos, puede que el uso de un sistema algorítmico en un entorno real contribuya a una pérdida de habilidades humana por parte de quienes utilizan los sistemas o que el acceso a una prestación a través de un sistema algorítmico contribuya al aumento de la brecha digital. Esto es, la búsqueda de la igualdad debe ir más allá de hablar de "malos datos" y "malos algoritmos" y de poner remedios a éstos[329], ahondando en los procesos por los que las instituciones, las normas y los sistemas (también algorítmicos) generan jerarquías sociales injustas[330].

3.2.1.2. Mitigación de sesgos en la toma de decisiones automatizada: un reto para la regulación europea de los sistemas de IA de alto riesgo

Ni desde un punto de vista técnico, ni normativo sería realista desarrollar o exigir sistemas de "riesgo cero". Con carácter general, la progresiva complejidad que los avances técnicos y científicos incorporan a nuestra vida social supone, a su vez, un aumento de los riesgos. No por ello renunciamos al uso de estos avances en determinadas actividades/profesiones

[327] Wachter, «Affinity profiling and discrimination by association in online behavioural advertising», 56.

[328] Adams-Prassl, Binns, y Kelly-Lyth, «Directly Discriminatory Algorithms», 32.

[329] Así lo explica McQuillan respecto de quienes tratan de desarrollar remedios anti-discriminatorios para los algoritmos: *While having the merit of trying to correct data science from a perspective that understands the technicalities of its operations, it is constrained by seeing data science as an external set of methods rather than as a broader social apparatus in Foucault's sense, that is 'a thoroughly heterogeneous ensemble consisting of discourses, institutions, architectural forms, regulatory decisions, laws, administrative measures, scientific statements, philosophical, moral and philanthropic propositions'*. Vid. McQuillan, «Data Science as Machinic Neoplatonism».

[330] Hoffmann, «Making Data Valuable: Political, Economic, and Conceptual Bases of Big Data», 911.

que comportan importantes riesgos: tráfico rodado, medicina, distribución industrial de productos para el consumo, industrias potencialmente contaminantes, etc.; si bien, el ordenamiento jurídico ocupa un papel determinante a la hora de establecer las reglas de conducta que debe regir el desarrollo y uso de estos avances, para que el riesgo introducido resulte aceptable a nivel colectivo[331].

De acuerdo con lo expuesto anteriormente, la forma en la que operan estos sistemas algorítmicos obliga a repensar las reglas del juego establecidas para la protección de distintos bienes jurídicos. Hemos visto que la normativa antidiscriminatoria es insuficiente, al menos en su forma actual, para proteger adecuadamente la igualdad material. Tampoco la normativa de protección de datos ofrece soluciones satisfactorias, y otras regulaciones parecen encontrar las mismas dificultades[332]. En este contexto y al hilo de las propuestas para la regulación de la IA, las instituciones europeas han manifestado una preocupación explícita por esta cuestión y han propuesto distintas estrategias para la mitigación de sesgos[333].

No obstante, las distintas estrategias para la mitigación de sesgos han encontrado contrapartidas -*tradeoffs*- en el plano normativo que han de ser discutidas y sopesadas adecuadamente.

La más discutida contrapartida es la elección entre el uso de datos de carácter sensible y la capacidad para entender la discriminación. La normativa de protección de datos está fundamentada en *instrumentos jurídicos de anti-clasificación*, esto es, normas que prohíben la consideración de las categorías especialmente sospechosas en los procesos de toma de decisión[334]. El principio de minimización de datos en el RGPD[335] es quizás el

[331] Desde una perspectiva del Derecho penal, Romeo Casabona, «El tipo del delito de acción imprudente», 133-34.

[332] Zuiderveen Borgesius, «Strengthening legal protection against discrimination by algorithms and artificial intelligence», 1581-82.

[333] La Comisión Europea consideró en su Libro Blanco sobre IA que deben adoptarse medidas razonables para velar porque el uso posterior de estos sistemas no genere discriminación, para lo cual considera la necesidad de adoptar medidas razonables como utilizar (1) conjuntos de datos que sean suficientemente representativos -requisitos sobre datos de entrenamiento- o (2) la necesidad de conservar documentación sobre metodologías de programación y entrenamiento con el fin de evitar resultados sesgados -requisitos sobre datos y registros de datos-. Vid. Comisión Europea, Libro Blanco sobre la inteligencia artificial.

[334] Soriano Arnanz, «La propuesta de Reglamento de inteligencia artificial de la UE y los sistemas de alto riesgo», 10.

[335] Art. 5(1) RGPD: *Los datos personales serán: (...) c) adecuados, pertinentes y limitados a lo necesario en relación con los fines para los que son tratados («minimización de datos»)*.

que mejor cristaliza este espíritu y revela, a mi juicio, la conexión histórica entre el derecho a la protección de datos con el derecho a la vida privada, y la tradicional concepción de la esfera más íntima del individuo como un espacio de especial protección frente a injerencias de terceros.

Sin embargo, el funcionamiento de los modelos de aprendizaje automático pone en duda la utilidad de estos instrumentos.

Los sistemas suelen entrenarse con datos que contienen sesgos implícitos; al mismo tiempo, los datos de entrenamiento pueden no contener referencias explícitas a la edad, el sexo u otros criterios. Así, la información decisiva no está presente y resulta imposible comprender si hay sesgo en los datos y, por consiguiente, también en el algoritmo, y si es posible ponerle remedio. Sin embargo, la inclusión de más datos, como la edad o el sexo, afecta al derecho a la intimidad y a la protección de datos. Especialmente en posibles casos de discriminación, a menudo sería necesario utilizar categorías especiales de datos personales, como los datos que revelan los orígenes raciales o étnicos, que están fuertemente protegidos por muchos regímenes de protección de datos[336]. Por lo tanto, es necesario sopesar la privacidad y la protección de datos con la aplicabilidad de la prohibición de discriminación en este contexto.

Así, la Comisión ha optado en su propuesta de Reglamento AIA por establecer una excepción al tratamiento de categorías especiales de datos por razones de interés público en su artículo 10(5) y que explica de la siguiente forma en el Considerando 44: «*Con el fin de proteger los derechos de terceros frente a la discriminación que podría provocar el sesgo de los sistemas de IA, los proveedores deben ser capaces de tratar también categorías especiales de datos personales, como cuestión de interés público esencial, para garantizar que el sesgo de los sistemas de IA de alto riesgo se vigile, detecte y corrija*».

Veremos si finalmente se mantiene dicha excepción en el texto aprobado por procedimiento legislativo ordinario. A pesar de que, dados los argumentos expuestos desde las ciencias de la computación[337], parece razonable defender que el tratamiento de

[336] Djeffal, «AI, Democracy and the Law», 269.

[337] Así los explican Wachter, Mittelstadt, y Russell: *The algorithmic fairness community in particular has developed methods which require special category data to detect and mitigate biases in training data and automated decisions. These communities are understandably increasingly calling for greater collection of*

estos datos sensibles permitiría un diseño y desarrollo de modelos más justos, resulta ingenuo pensar que unos resultados más justos e igualitarios provendrán de, simplemente, recolectar más datos[338].

La complejidad de los sesgos obliga a aplicar distintas estrategias sobre diferentes fases de la toma de decisiones, asistida o no por máquinas. También a discutir qué distintos enfoques normativos pueden ser más efectivos en cada etapa para garantizar el cumplimiento normativo[339].

Para hacer frente a esta compleja tarea es necesario repensar las posibilidades sobre la "eliminación" de sesgos en entornos complejos, especialmente a los argumentos que apelan a una IA "libre de sesgos". Si estas tecnologías operan indispensablemente en un entorno social o ecosistema determinado por personas y estructuras sesgadas, incluso en el caso de que pudiese desarrollarse una tecnología "libre de sesgos", su utilidad resultaría cuestionable[340]. Es por ello que las estrategias, normativas o no, que tengan por objetivo la mitigación de sesgos, no deberían limitarse a la aparición y reproducción de sesgos en los datos de entrada y salida de un programa informático, sino extenderse a las personas y estructuras que forman parte del proceso de toma de decisiones[341]. De otro modo, una

special category data to facilitate discrimination detection and legal proceedings. Wachter, Mittelstadt, y Russell, «Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI», 35.

338 Wachter, Mittelstadt, y Russell, «Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law», 45-47.

339 Xenidis y Senden realizan una comparativa entre un enfoque basado en derechos individuales y otro basado en la vigilancia de una autoridad pública, que desde luego resultan compatibles, mientras que Zuiderveen Borgesius aboga por un enfoque sectorial teniendo en cuenta que no puede evaluarse la equidad de un sistema algorítmico en abstracto. Vid. Xenidis y Senden, «EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination», 23-29; Zuiderveen Borgesius, «Strengthening legal protection against discrimination by algorithms and artificial intelligence», 1484-85.

340 En este sentido, Danks y London concluyen: *More generally, we need to think about algorithmic bias (with respect to various norms) in terms of the whole system, including the consumer human or machine of the algorithm output. The "ecosystem" around an algorithm contains many opportunities for both the introduction of bias, and also the injection of compensatory biases to minimize the harms (if any) done by the algorithmic bias*. Danks y London, «Algorithmic Bias in Autonomous Systems», 4696.

341 Es por ello que Lin et al. abogan por una perspectiva en la que la mitigación de sesgos responda a la complementariedad humano-máquina y a la búsqueda de la mejor interacción posible. Vid. Lin, Hung, y Huang, «Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias».

tecnología "libre de sesgos" podría ser utilizada para camuflar el fracaso de las instituciones a la hora de paliar las relaciones de desigualdad existentes en el sistema[342].

4. Opacidad en la toma de decisiones automatizada basada en la elaboración de perfiles

En este capítulo, a partir de la definición de las distintas fases que podemos encontrar en la toma de decisiones algorítmica, se han analizado dos aspectos característicos y determinantes de estos sistemas: la participación humana en la implementación de los sistemas y la producción y reproducción de sesgos en los mismos. Si bien, para poder completar este marco teórico de la toma de decisiones automatizada basada en la elaboración de perfiles, es necesario acercarse a la cuestión que más debate ha suscitado en la literatura: el funcionamiento oscuro u opaco de los sistemas algorítmicos.

Una opacidad que se ha convertido en norma para las instituciones privadas y que incluso las instituciones públicas han adoptado igualmente[343], y que fue ilustrada con la metáfora ampliamente conocida gracias a la obra de Pasquale: *The Black Box Society*.

Los efectos nocivos de la opacidad han sido ampliamente documentados; puede impedir que los desarrolladores intervengan para mejorar rápida y sistemáticamente el rendimiento del sistema o corregir sus errores[344]; los usuarios finales son menos propensos a confiar y ceder el control a máquinas cuyo funcionamiento no entienden[345]; dificulta el cumplimiento de derechos individuales como el de información y acceso al RGPD[346]; y puede funcionar como un velo para ocultar la normativa que se elude -

[342] Krupiy, «A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective», 21.

[343] de Laat, «Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?», 2. En España es destacable el caso del software BOSCO puesto en marcha en 2017 para el cálculo del bono social de electricidad y que determina quién tiene acceso a dichas ayudas y quién no. En 2019 la plataforma Civio, con la asistencia letrada de Javier de la Cueva González-Cotera, interpuso demanda en primera instancia ante la jurisdicción contenciosa contra la negativa del Ministerio para la Transición Ecológica a liberar el código fuente del programa BOSCO. La primera página de la demanda está disponible en el siguiente enlace: https://civio.app.box.com/s/6idsfsw1ny1odw0sffjrj6gwtx9pru5c

[344] Hohman et al., «Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers».

[345] Zednik, «Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence», 266.

[346] Goodman y Flaxman, «European Union regulations on algorithmic decision-making and a "right to explanation"».

también para las autoridades reguladoras-, la manipulación de los consumidores y/o los patrones de discriminación[347].

Cabe preguntarse si el hecho de que para las personas usuarias de estos sistemas sea prácticamente imposible comprender cómo un modelo de estas características alcanza un determinado resultado o dato de salida, requiere necesariamente que los modelos algorítmicos basados en IA sean una amenaza para el Estado de Derecho y el cumplimiento normativo.

Cuando se hace referencia a que estos sistemas son *cajas negras*, nos referimos con carácter general a que su capacidad predictiva reside, precisamente, en establecer correlaciones tan sumamente complejas que sus resultados no son interpretables, en el sentido de inteligibles, a partir de los datos –inputs– que se aportan al modelo para el ser humano, ni siquiera para quienes los diseñan[348]. No obstante, esta no es la única forma de opacidad que provoca un efecto caja negra en la implementación y uso de estos modelos, tampoco hay razones para concluir que es necesariamente la más perjudicial.

En este apartado trataré de aportar una sistematización de las distintas clases y niveles de opacidad que pueden darse en estos modelos. Posteriormente, introduciré la relación entre estas clases de opacidad y el concepto normativo de transparencia como remedio a los efectos nocivos de la opacidad.

4.1. Clases de opacidad

Tal y como se ha señalado arriba, la opacidad es polisémica en la literatura y no se refiere únicamente a la incapacidad humana para interpretar los resultados de determinados modelos algorítmicos. Aunque se ha tomado como referencia el conocido trabajo de Burrell y las clases de opacidad que ella distinguió[349], se han añadido otras dos formas de

[347] Burrell, «How the machine 'thinks': Understanding opacity in machine learning algorithms», 4.

[348] Selbst y Barocas, «The Intuitive Appeal of Explainable Machines», 1094.

[349] Esto es*: (1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy and (3) opacity as the way algorithms operate at the scale of application*, aquí opacidad intencionada o deliberada, opacidad código y opacidad inherente, respectivamente. Vid. Burrell, «How the machine 'thinks': Understanding opacity in machine learning algorithms».

opacidad halladas en la literatura y que son de relevancia para el análisis normativo de la cuestión[350].

La esencia del concepto de "caja negra" proviene de que dicha "caja" -pongamos sistema automatizado o algorítmico- genera una acción o resultado al tiempo que su carácter opaco impide el acceso al proceso subyacente. Dice Andrejevic que este concepto revela una tensión metafórica con la imagen de "luz" o transparencia en el corazón de la concepción ilustrada del conocimiento: que es susceptible de explicación racional y, por tanto, compartible y explicable[351]. De ahí también que la transparencia haya sido propuesta como remedio normativo, como veremos más adelante.

En la sistematización aquí realizada sobre las clases de opacidad, veremos que la mayoría de ellas son formas de opacidad ya conocidas por el Derecho y cuentan con una categorización ya establecida, y en este sentido la única "novedad" es la aparición de la opacidad inherente a los modelos algorítmicos que son objeto de estudio en esta investigación. Ahora bien, y sin perjuicio de que las respuestas jurídicas particulares a cada clase de opacidad puedan variar, esta sistematización resulta necesaria a la vista de la confusión producida por la laxitud con la que se han utilizado términos como "opacidad" o "cajas negras" en la literatura. Tal y como se ilustra a continuación, las clases y subclases de opacidad que se definen pueden solaparse parcialmente y combinarse de distintas maneras, visualizar este solapamiento es útil para atender a los efectos que puede producir la opacidad en sus distintas formas, así como para identificar las respuestas jurídicas adecuadas en cada caso a la hora de mitigar dichos efectos:

4.1.1. Opacidad inherente al modelo

Esta clase de opacidad es, sin lugar a duda, la más discutida en la literatura y deriva de que las funciones y correlaciones utilizadas por muchos de los algoritmos predictivos basados en aprendizaje automático para la toma de decisiones[352], pueden ser demasiado

[350] Esto es, la opacidad normativa y la opacidad procedimental.

[351] Andrejevic, «Shareable and un-sharable knowledge», 1.

[352] En oposición a los algoritmos utilizados en la programación clásica, que no presentan esta dificultad, Walmsley explica que las áreas de la IA contemporánea que más rápido avanzan son aquellas en las que los humanos están en desventaja epistémica a la hora de entender los sistemas que hemos construido: *At the intersection of machine learning, big data, and algorithm appreciation, we have a situation, where we do not fully understand the machines, they're faster, more powerful and more complex than us, but we trust them preferentially nonetheless*. Walmsley, «Artificial intelligence and the value of transparency», 586-87.

complejas para que las comprendan los seres humanos, y en este sentido, no es interpretable para los humanos cómo el algoritmo ha llegado a una determinada decisión a partir de determinados datos de entrada.

Esta clase de opacidad despierta una preocupación si cabe mayor que las demás, dado que se proyecta incluso sobre quienes desarrollan estos modelos, es decir, sobre sus propias creaciones[353]. No puede obviarse que esta clase de opacidad responde, en cierto sentido, a una decisión deliberada adoptada en el desarrollo del modelo. Dado que no todos los modelos algorítmicos, ni siquiera todos los modelos de aprendizaje automático padecen esta falta de interpretabilidad humana. Cierto es que una contrapartida o *trade-off* habitualmente discutida en la literatura es que los modelos más complejos obtienen una mayor precisión que los modelos más interpretables, con lo cual, hay que sopesar las ventajas de modelos más interpretables -control y rendición de cuentas- frente a las de modelos menos interpretables -precisión-[354].

En cualquier caso, equipos multidisciplinares están trabajando en soluciones técnicas que permitan paliar esta clase de opacidad en el campo de la interpretabilidad y explicabilidad algorítmica[355]. Por lo general, aunque hay muchas clases de modelos dirigidos también a distintos usuarios del ciclo de vida de un sistema, el objetivo de estos sistemas es ser de utilidad a los y las profesionales a la hora de interpretar las correlaciones elaboradas por el sistema y muy especialmente cómo éste podría fallar[356].

Desde una perspectiva jurídica, la explicabilidad algorítmica tiene por incentivo principal garantizar que las explicaciones de las decisiones algorítmicas permitan que la persona afectada comprenda las razones que subyacen a la decisión y evitar decisiones discriminatorias o que no se ajusten a la Ley, por lo general favoreciendo la contestabilidad de la decisión. También para que los responsables humanos que utilicen algún tipo de sistema de apoyo a la toma de decisiones puedan considerarse responsables

[353] Carabantes, «Black-box artificial intelligence: an epistemological and critical analysis», 2.

[354] de Laat, «Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?», 15.

[355] Acerca de las diferencias entre las nociones de "interpretabilidad" y "explicabilidad" y su eventual relevancia vid. Kiseleva, Kotzinos, y De Hert, «Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations», 6-7.

[356] Sobre el estado del arte de esta cuestión puede consultarse: Gilpin et al., «Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning», 80-89. Para un acercamiento más filosófico vid. Mittelstadt, Russell, y Wachter, «Explaining Explanations in AI»..

de las decisiones adoptadas finalmente[357]. En cambio, para la ciencia de la computación los incentivos abarcan consideraciones éticas y técnicas que trascienden dicho objetivo normativo[358].

Tal y como explica Brkan, la viabilidad de las explicaciones algorítmicas depende de (1) qué clase de explicaciones queremos/debemos aportar (por ejemplo, diferenciamos entre explicaciones locales que tratan de identificar los principales factores que influyen en el funcionamiento de un modelo, y explicaciones contrafactuales que evalúan el peso de diferentes factores en la toma de decisiones particular), (2) en qué momento queremos aportarlas (ex ante o ex post), (3) qué modelos son más adecuados para proporcionar unas u otras explicaciones y (4) del mejor método técnico disponible para la generación de las explicaciones[359].

Y es que esta clase de opacidad no requiere que tengamos que renunciar a toda explicación acerca del sistema. Andrejevic explica que, al menos en principio, siempre será posible evaluar el impacto global del sistema, aunque no podamos hacer ingeniería inversa de los procesos que lo han provocado. Eso sí, no deben infraestimarse los costes económicos y las relaciones de poder que gobiernan esta posibilidad de evaluar el impacto global de un sistema[360].

4.1.2. Opacidad intencionada o deliberada

[357] Cabitza, Campagner, y Ciucci, «New Frontiers in Explainable AI: Understanding the GI to Interpret the GO BT - Machine Learning and Knowledge Extraction», 42. Muy relevante también la siguiente distinción que trazan en la interacción entre el ser humano y la IA; entre el derecho a la explicación, es decir, a que los usuarios reciban indicaciones por parte del sistema de IA que les hagan creer satisfactoriamente que han entendido por qué el sistema de apoyo a la toma de decisiones les ha dado un determinado resultado; y la obligación de interpretación, es decir, que los usuarios tengan que adoptar una actitud activa para recoger e interpretar estas indicaciones. Concluyen, que la defensa de una IA explicable no debe disminuir la responsabilidad de los responsables de la toma de decisiones.

[358] Brkan y Bonnet, «Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas», 19.

[359] Brkan y Bonnet, 47.

[360] Así dice: *With enough data and the power to analyze it, we can determine whether, for example, a system's outcomes exhibit a particular pattern (of discrimination, for example), even if we cannot open the box to explain why. In many cases, however, the data will be hard to collect or proprietary, and the cost of collecting it and analyzing it prohibitive*. Andrejevic, «Shareable and un-sharable knowledge», 4.) También Boix incide sobre estos costes, aunque advierte: *Frente a esta situación, ha de ser señalado que las garantías jurídicas no son necesariamente baratas. Es más, no lo han sido nunca. Ni las que aquí se proponen, ni las tradicionales. Y es preciso recordar también que, aun costosas, su no respeto suele conllevar también costes considerables a medio y largo plazo*. Boix, «Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones», 259.

La opacidad deliberada hace referencia a la decisión por parte de quien desarrolla o implementa un modelo automatizado de, por razones muy diversas, ocultar todo o parte del sistema a terceros implicados en la toma de decisiones.

Pueden definirse dos niveles de opacidad en este marco: uno primario en el que la persona interesada o afectada desconoce que el proceso decisorio se realiza por medio de un tratamiento automatizado de datos[361]; y uno secundario, donde la interesada sí sabe que la decisión está basada en un tratamiento automatizado, sin embargo, no se le permite conocer el funcionamiento interno del proceso decisorio automatizado. Ambos niveles de opacidad deliberada pueden resultar lesivos para quienes soportan las decisiones finales de los sistemas automatizados, en el primer caso porque los sistemas operan de forma invisible y en el segundo porque operan de forma incomprensible[362].

Ahora bien, los términos intencional o deliberado no deben llevarnos a una concepción necesariamente peyorativa de estas decisiones; es cierto que, tal y como afirma Pasquale en *The Black Box Society,* en el ámbito financiero estas decisiones responderían incluso a un intento de evadir la regulación o confundir a las instituciones reguladoras[363]; sin embargo, ello no quiere decir que las mismas no puedan responder también a la protección de intereses o derechos propios o de terceros, por ejemplo, la protección de datos personales de terceros protegidos por la normativa de protección de datos que podrían resultar expuestos con la revelación del modelo, o la seguridad del sistema para evitar que

[361] Los efectos de esta clase de opacidad primaria se infraestiman en ocasiones, cuando ya son millones de decisiones por segundo las que se producen a nuestro alrededor basadas en estos modelos algorítmicos, y la tendencia a delegar estas decisiones y clasificaciones invisibles no hace más que aumentar. Carabantes, «Black-box artificial intelligence: an epistemological and critical analysis», 2.

[362] Estos dos niveles se explicaban por el Parlamento Europeo en el anexo a su Resolución de 20 de octubre de 2020, con recomendaciones destinadas a la Comisión sobre un marco de los aspectos éticos de la inteligencia artificial, la robótica y las tecnologías conexas (2020/2012(INL)), de la siguiente manera en el Considerando 21: *Para garantizar la transparencia y la rendición de cuentas, se debe informar a los ciudadanos siempre que un sistema utilice inteligencia artificial, siempre que los sistemas de inteligencia artificial personalicen un producto o un servicio para sus usuarios, así como de si pueden desactivar o limitar la personalización, y siempre que se enfrenten a una tecnología de toma de decisiones automatizada. Además, las medidas de transparencia deben ir acompañadas, siempre que sea técnicamente posible, de explicaciones claras y comprensibles sobre los datos utilizados y el algoritmo, así como sobre su finalidad, sus resultados y sus riesgos potenciales*.

[363] Pasquale, *The Black Box Society*, 2.

los modelos sean objeto de fraude -habitualmente conocido como *game the system* en literatura anglófona[364]-, o mantener la competitividad entre distintos actores privados[365].

No obstante, muchos de estos fines legítimos son utilizados de forma abusiva por las corporaciones privadas e instituciones públicas, como treta para ocultar que sus algoritmos no cumplen con la legalidad vigente o que son éticamente cuestionables[366]. También otras formas de opacidad han sido utilizadas en el discurso público para enmascarar una opacidad deliberada, así McQuillan resalta que la opacidad algorítmica ha sido, en gran parte, consecuencia de los "altos muros" del secreto comercial y no solo de la tendencia de determinados modelos a la opacidad por naturaleza o inherente[367].

4.1.3. Opacidad código

Esta clase de opacidad se produce dado que, en la actualidad, escribir y leer código fuente y el diseño de algoritmos es una habilidad especializada, que continúa siendo inaccesible para la mayoría de la población[368].

Hay quien podría señalar que la opacidad código disminuye la utilidad de remedios normativos como la publicación del código fuente dado que no sería accesible para la ciudadanía. No obstante, este argumento resulta endeble y paternalista. Boix lo expresa con la siguiente analogía: «*Las normas jurídicas tradicionales son también, o pueden serlo en demasiadas ocasiones, muy opacas para los no especialistas, pero ello no es razón para que no se publiquen sino, antes al contrario, una situación que hace si cabe más necesaria la total transparencia, como medio de garantizar que al menos puedan estar a disposición potencial de cualquier posible especialista en la materia que pueda*

[364] Este es uno de los argumentos del gobierno neerlandés para denegar el acceso a su modelo de riesgo para investigar el fraude social en el caso SyRI resuelto por el Tribunal de Distrito de La Haya. El Estado se niega a facilitar esta información alegando que para luchar contra el fraude es fundamental obtener «datos de calidad», entendidos como aquellos que vigilan el comportamiento de los ciudadanos sin que éstos lo sepan o sin que sepan en qué se fijan. Como se refleja en el párrafo 6.49, el Estado decide no dar cierta información sobre el funcionamiento de SyRI argumentando que los ciudadanos podrían en tal caso ajustar su comportamiento a los parámetros del algoritmo. Dicho de otro modo, si los ciudadanos conocen los datos en que SyRI se fija y ajustan su comportamiento, el conocimiento obtenido de vigilarlos no sería igualmente útil en la medida en que la información obtenida no sería de la misma calidad. Vid. par. 6.49

[365] de Laat, «Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?», 10-14.

[366] Vid. Pasquale, *The Black Box Society*.

[367] McQuillan, «Data Science as Machinic Neoplatonism», 257.

[368] Burrell, «How the machine ‘thinks’: Understanding opacity in machine learning algorithms», 4.

existir en el mundo con capacidad para entender y comprender tanto el contenido de la programación normativa como sus implicaciones, así como detectar posibles errores en la misma»[369]. Me permito añadir que esta dificultad aparece en casi todos los ámbitos, casi ningún paciente sabe interpretar un electroencefalograma y ello no es excusa para impedir el acceso a sus resultados, por ejemplo, para solicitar una segunda opinión médica conforme a la normativa sanitaria.

Además, la educación/formación ha sido propuesta como solución parcial[370], es decir, ampliar y generalizar los conocimientos sobre el código y las habilidades computacionales para mitigar el problema de una clase homogénea y elitista de personas técnicas que toman decisiones consecuentes que no pueden ser evaluadas fácilmente por el resto[371]. Resulta inevitable traer aquí el símil de la alfabetización como condición necesaria, aunque no suficiente, para la democracia[372]. Desde esta perspectiva educativa, se ha llamado además la atención sobre el papel que juega la ausencia de diversidad en el ámbito STEM (ciencia, tecnología, ingeniería y matemáticas) en el contexto y desarrollo de estos sistemas[373]. Así como sobre el papel que el periodismo pudiera desempeñar en la información a un público general sobre las decisiones algorítmicas que tienen efectos sobre sus vidas[374].

4.1.4. Opacidad normativa

[369] Boix, «Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones», 257.

[370] La Carta de Derechos Digitales del Gobierno de España incluye un derecho a la educación, cuyo apartado primero dice así: *El sistema educativo debe tender a la plena inserción de la comunidad educativa en la sociedad digital y un aprendizaje del uso de los medios digitales dirigido a una transformación digital de la sociedad centrada en el ser humano. Esta misión se inspirará en los valores de respeto de la dignidad humana con garantía de los derechos fundamentales y los valores constitucionales. Estos principios informarán cualesquiera otras actividades formativas promovidas por los poderes públicos*. Vid. Gobierno de España, Carta de Derechos Digitales, Plan de recuperación, transformación y resiliencia, 2021. Texto completo disponible aquí: https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf

[371] Burrell, «How the machine 'thinks': Understanding opacity in machine learning algorithms», 10.

[372] Vid. Rodas-García, «Alfabetizar para la democracia».

[373] El despido de Google de la investigadora Timnit Gebru ilustra perfectamente parte de esta problemática, que tiene sus efectos también sobre quién y cómo ejerce el poder en este ámbito. Vid. Pérez Colomé, «Por qué el despido de una investigadora negra de Google se ha convertido en un escándalo global». Disponible en: https://elpais.com/tecnologia/2020-12-12/por-que-el-despido-de-una-investigadora-negra-de-google-se-ha-convertido-en-un-escandalo-global.html

[374] Carabantes, «Black-box artificial intelligence: an epistemological and critical analysis», 4.

Al abordar la opacidad deliberada o intencionada, se ha hecho referencia a que el fundamento de esta clase de opacidad puede responder a distintos motivos, entre otros: razones de secreto industrial y ventajas competitivas, privacidad sobre los datos utilizados por el modelo o seguridad y protección frente a ataques adversarios o la utilización fraudulenta del sistema. Así, en no pocas ocasiones, el acceso a la información sobre el modelo tiene un fundamento normativo que funciona, a efectos prácticos, como una forma de opacidad normativa. Puede que la aplicación de la norma sea imperativa -revelar datos de terceros sin una base legitimadora para hacerlo estaría prohibido en todo caso- o que sirva de fundamento a una opacidad deliberada -en principio una empresa que desarrolla un algoritmo puede decidir si dar acceso libre a su código o protegerlo por razones de secreto industrial-.

Como veremos a continuación, en muchos casos la jurisprudencia ya se ha enfrentado a esta clase de opacidad y a la ponderación necesaria ante normas de transparencia que, en sentido contrario, obligan a revelar información.

Un aspecto muy discutido en la doctrina es la posible opacidad derivada derechos que permiten proteger descubrimientos e invenciones, esto es, la normativa acerca de derechos de propiedad intelectual e industrial y de secretos comerciales e industriales. Según Castillo Parrilla, en nuestro ordenamiento la protección jurídica de los algoritmos proviene de su naturaleza como descubrimientos –en tanto el código es una expresión matemática– y no como invenciones. Lo cual deriva en una situación de doble exclusión jurídica en la UE: tanto del régimen de propiedad industrial como del régimen de propiedad intelectual[375], y ello supone que la protección jurídica de los algoritmos se obtenga a partir de la vía residual de los secretos industriales[376], actualmente en nuestro

[375] Castillo Parrilla, «El turismo en la economía de los datos y la economía de plataformas en la UE», 130.

[376] La Directiva 943/2016 de 8 de junio de 2016, relativa a la protección de los conocimientos técnicos y la información empresarial no divulgados (secretos comerciales) contra su obtención, utilización y revelación ilícitas, define los secretos comerciales en el artículo 2(1) como la información que reúna todos los requisitos siguientes: *a) ser secreta en el sentido de no ser, en su conjunto o en la configuración y reunión precisas de sus componentes, generalmente conocida por las personas pertenecientes a los círculos en que normalmente se utilice el tipo de información en cuestión, ni fácilmente accesible para estas; b) tener un valor comercial por su carácter secreto; c) haber sido objeto de medidas razonables, en las circunstancias del caso, para mantenerla secreta, tomadas por la persona que legítimamente ejerza su control*;

ordenamiento regulados por la Directiva 943/2016, de 8 de junio, de secretos comerciales y la Ley 1/2019, de 20 de febrero, de Secretos Empresariales[377].

Ello no quiere decir que todos los elementos que integran la elaboración de perfiles para la toma de decisiones automatizada no sean susceptibles de ser protegidos por derechos de propiedad intelectual, y en este sentido es necesario distinguir los algoritmos del código fuente -*source code*- y del programa informático -*software*-, que pueden ser objeto de protección por estos derechos, aunque ello no incluya al algoritmo que forma parte de éstos[378]. En el conocido caso *Loomis* sobre el sistema COMPAS -utilizado para valorar la peligrosidad de un individuo en un proceso penal-, el Tribunal Supremo de los Estados Unidos denegó el acceso al sistema -sobre el fundamento del derecho al debido proceso- imponiendo las razones de secreto comercial[379].

El derecho a la vida privada y a la protección de datos actúa también como una forma de opacidad normativa. A fin de cuentas, tanto en el entrenamiento de un modelo como en la implementación del mismo, la utilización de datos personales que cuentan con protección normativa es muy habitual. Comprender el funcionamiento del modelo depende en gran medida de la clase de acceso que pueda tenerse a dichos datos, por ello, la aplicación de la normativa de protección de datos puede obstaculizar dicha comprensión. El RGPD prevé esta colisión para el ejercicio de los derechos de información y acceso de la persona interesada.

---

[377] Castillo Parrilla, «El turismo en la economía de los datos y la economía de plataformas en la UE», 132. Lo que genera efectos perniciosos a su juicio: *Si la única vía expedita para su protección es la de los secretos comerciales, parece esperable que quienes invierten dinero y esfuerzo en su desarrollo no sean muy proclives (atendiendo exclusivamente a argumentos jurídicos relacionados con la protección de su inversión en este punto) a revelar su funcionamiento o facilitar su transparencia*.

[378] Brkan y Bonnet describen perfectamente esta diferenciación en Brkan y Bonnet, «Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas», 45.

[379] En contra de la ponderación del Tribunal, Romeo Casabona expone: *No encaja muy bien, sin embargo, esta preferencia del tribunal de proteger los derechos e intereses de la empresa privada frente a los derechos procesales del condenado, que se ven así postergados; en concreto, haciendo una traslación hipotética con los principios de nuestro ordenamiento jurídico, como son el de proporcionalidad y con el derecho fundamental a la tutela judicial efectiva frente a los derechos de carácter económico, que gozan de un rango constitucional inferior. Al parecer sí es más afín a la mentalidad jurídica estadounidense el criterio opuesto, pero no deja de constituir una inversión del rango de derechos en conflicto*. Romeo Casabona, «Riesgo, procedimientos actuariales basados en inteligencia artificial y medidas de seguridad», 53.

En los casos resueltos por el Tribunal de Distrito de Ámsterdam acerca de las disputas entre las compañías Uber y Ola y varios de sus *drivers*[380], el tribunal aplica esta clase de limitaciones al derecho de acceso y que se ubican en la misma norma. Así, aplicó el artículo 15(4) RGPD para proteger los derechos de protección de datos de terceros -clientes-, que dice: «*El derecho a obtener copia mencionado en el apartado 3 no afectará negativamente a los derechos y libertades de otros*». No obstante, para la aplicación de esta limitación ha de tenerse en cuenta que el Considerando 63 desarrolla la misma, añadiendo que dicha limitación no debe tener como resultado la negativa a prestar toda la información al interesado. Una solución que efectivamente aplica el Tribunal, y que puede resultar de gran utilidad en cualquier contexto para la protección de derechos de protección de datos de terceros, es la anonimización o seudonimización de la información[381].

4.1.5. Opacidad procedimental

Mientras que en los anteriores supuestos me he centrado en las formas de opacidad que se proyectan directamente de forma "externa" (por quien desarrolla o implementa el modelo respecto de la persona interesada u objeto de las decisiones o de quien fiscaliza el modelo), en esta última categoría me gustaría centrarme en la opacidad que se proyecta a nivel "interno", es decir, entre quien desarrolla y quien implementa el modelo (lo cual, a su vez, puede repercutir en opacidad proyectada a nivel externo). Esta clasificación la ilustra perfectamente Kiseleva en el ámbito sanitario, mientras que la opacidad externa tiene su proyección directa sobre el derecho a la autonomía de los pacientes, la opacidad interna afecta a la toma de decisiones clínica por parte de los profesionales sanitarios[382].

---

[380] Vid. referencias de estas sentencias en Capítulo 2. Toma de decisiones automatizada basada en la elaboración de perfiles en el RGPD: el artículo 22 en la unidad de cuidados intensivos..

[381] Vid. análisis más extenso en Lazcoz, «Automated decision-making under Amsterdam's District Court judgements: Drivers v. Uber and Ola». Por otro lado, y con un fundamento más polémico, el tribunal limita el derecho de acceso de los *drivers* con fundamento directo en el Considerando 63 RGPD: Si trata una gran cantidad de información relativa al interesado, el responsable del tratamiento debe estar facultado para solicitar que, antes de facilitarse la información, el interesado especifique la información o actividades de tratamiento a que se refiere la solicitud. Tal y como he mantenido en trabajos previos, trasladar al interesado la carga de especificar la información sobre el tratamiento cuánto más compleja es la cantidad de información tratada por el responsable parece contradictorio, más cuando el responsable del tratamiento no proporcione información concisa, transparente, inteligible y de fácil acceso con arreglo al artículo 12 del RGPD.

[382] Una vez más, esta opacidad interna o procedimental, afecta en último término a la posibilidad de informar adecuadamente a los pacientes, en los términos requeridos por el consentimiento informado. Vid.

La opacidad procedimental entre quien desarrolla e implementa el modelo puede reproducirse en cualquiera de las clases de opacidad antes descritas cuando éstas dependen del desarrollador. Es decir, como es evidente, la opacidad deliberada en su primer nivel -en el que la persona interesada o afectada desconoce que el proceso decisorio se realiza por medio de un tratamiento automatizado de datos- no puede producir una opacidad procedimental, dado que la decisión de ocultar el sistema automatizado de la toma de decisiones es responsabilidad de quien implementa el modelo. Al contrario, la opacidad deliberada en su segundo nivel -pongamos que el desarrollador protege bajo un fundamento normativo (secreto industrial) el funcionamiento interno del sistema automatizado- puede producir una opacidad procedimental.

Esta clase de opacidad puede ser especialmente relevante en contextos en los que la intervención o supervisión humana se establece como salvaguarda del ordenamiento jurídico, considerando que dicha intervención es fundamento para no levantar otras formas de opacidad externa. De hecho, las dificultades para garantizar una supervisión humana "significativa" se ven exacerbadas en este contexto de opacidad[383].

4.2. Transparencia como principio o fin normativo

La discusión jurídica sobre cómo combatir los efectos nocivos de la opacidad ha llevado a buscar la solución que parece más sencilla: permitir la entrada de luz en estos modelos[384]. De ahí que la transparencia haya sido el principio normativo sobre el que han discurrido la mayoría de los debates doctrinales en este contexto.

En cuanto al término "transparencia", desde un punto de vista técnico podría entenderse como la habilidad para observar los detalles internos de un sistema, de tener acceso al

Kiseleva, «AI as a Medical Device: Is it Enough to Ensure Performance Transparency and Accountability?».

[383] Noto La Diega, «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information», 19. En este sentido, en el mencionado caso Loomis sobre el sistema COMPAS, Romeo Casabona señala la incongruencia del Tribunal al insistir en este sistema como un medio de prueba más a disposición del decisor humano, obviando las -reconocidas- limitaciones del sistema que se proyectan hacia éste, y no únicamente hacia el sujeto al que afecta la resolución judicial como tal. Vid. Romeo Casabona, «Riesgo, procedimientos actuariales basados en inteligencia artificial y medidas de seguridad», 177.

[384] de Laat, «Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?», 3.

código del algoritmo, por ejemplo[385]; mientras que desde un punto de vista ético-jurídico la transparencia adquiere o debería adquirir un significado más profundo. Felzmann et al. explican que muchos de los debates se basan en una perspectiva puramente informativa de la transparencia, como la transferencia de información de un agente a otro. No obstante, y aunque éste sea un componente esencial de cualquier comprensión de la transparencia, ignora la profunda carga de valores de la transparencia en la agencia individual y en las prácticas relacionales y sistémicas. La transparencia no puede entenderse sin prestar atención a dichos valores y a las funciones sociales asociadas a esta transferencia de información[386].

La preocupación por la transparencia suele obedecer a una determinada cadena de lógica: la observación produce conocimientos que permiten gobernar los sistemas y exigir una debida responsabilidad a quienes los desarrollan, implementan y usan[387]. El control o la autonomía y la transparencia no están tan directamente relacionados como pudiera parecer. Por ejemplo, si quien resulta destinatario de la transferencia de información no puede aprovechar la misma porque sigue siendo de difícil acceso, se presenta de forma compleja u opaca, o no tiene una alternativa significativa a la decisión sobre la que se le informa, no hay un aumento de la autonomía y del control asociado a dicha práctica de transparencia[388]. Por todo ello, la aplicación de normas de transparencia sobre sistemas algorítmicos obliga a una evaluación más allá de un quién revela, qué y a quién[389].

En el ámbito de la privacidad y la protección de datos, esta perspectiva normativa basada en la transparencia para la gobernanza de la opacidad algorítmica ha ocupado un lugar central. Durante la tramitación legislativa del RGPD, en su Dictamen 3/2015, el

[385] Yeung y Weller, «How is 'transparency' understood by legal scholars and the machine learning community?»

[386] Felzmann et al., «Towards Transparency by Design for Artificial Intelligence», 3336.

[387] Ananny y Crawford, «Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability», 974.

[388] Felzmann et al., «Towards Transparency by Design for Artificial Intelligence», 3341.

[389] En este sentido son muy ilustrativas las distintas tipologías de transparencia que recogen Ananny y Crawford de trabajos previos como de Fox y Heald y que distingue entre transparencia difusa y clara según el grado de fiabilidad de la información que traslada; transparencia que genera una responsabilidad blanda o dura en función del poder de respuesta jurídica que conlleva; transparencia hacia arriba, hacia abajo, hacia dentro y hacia fuera, según a quienes obliga y quién es el destinatario de la información; transparencia como un evento, por definir los objetos de la transparencia, o como un proceso, por definir las condiciones de visibilidad; y transparencia retrospectiva frente a transparencia en tiempo real. Vid. Ananny y Crawford, «Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability», 976.

Supervisor Europeo de Protección de Datos (SEPD) expresó de este modo la centralidad de la transparencia: «*El problema no es (…) la práctica de la elaboración de perfiles, sino, más bien, la falta de información adecuada sobre la lógica algorítmica a partir de la que se desarrollan tales perfiles y que repercute en el interesado*»[390]. De este modo, el principio de transparencia (art. 5(1) RGPD) se tradujo en obligaciones para los responsables del tratamiento de proveer de información a las personas interesadas, y en particular en derechos de información y acceso a *información significativa sobre la lógica aplicada* en la toma de decisiones automatizadas, así como la importancia y las consecuencias previstas (arts. 13(2)(f), 14(2)(g) y 15(1)(h) RGPD). A su vez, dando lugar también a un intenso debate doctrinal acerca del alcance de estos derechos[391].

Desde el Derecho Administrativo se ha discutido si debe o no reconocerse un carácter normativo a los algoritmos -especialmente cuando se reconocen efectos directos a los resultados algorítmicos- y, por ende, deben aplicarse las exigencias de publicidad normativa a estos algoritmos abriendo el código fuente de los mismos o, por el contrario, son de aplicación únicamente las normas de transparencia y protección de datos[392]. Las regulaciones europea y española parecen orientarse claramente a esta última posición [393], lo cual ha llevado prácticamente a las mismas limitaciones que la doctrina ha encontrado en la aplicación del derecho de protección de datos.

Parece razonable defender la existencia de un doble estándar de transparencia entre la toma de decisiones humana y algorítmica. Es decir, para la toma de decisiones humana no exigimos conocer el proceso interno que lleva a la adopción de una determinada decisión y, no obstante, establecemos procedimientos alternativos para confiar y fiscalizar estas decisiones, mientras que la exigencia de hacer público este proceso interno está muy

[390] Supervisor Europeo de Protección de Datos (SEPD), «Dictamen 3/2015 -La gran oportunidad de Europa. Recomendaciones del SEPD sobre las opciones de la UE en cuanto a la reforma de la protección de datos-», 10.

[391] Esta cuestión será abordada en profundidad más tarde. Vid. Apartado 3.2.2. Derecho a la información para las decisiones basadas únicamente en el tratamiento automatizado, ¿derecho a una explicación? Capítulo 2. Toma de decisiones automatizada basada en la elaboración de perfiles en el RGPD: el artículo 22 en la unidad de cuidados intensivos.

[392] Boix defiende con sugerentes argumentos la posición minoritaria en España (los algoritmos son reglamentos) y ha dado lugar a un interesante debate como atestigua la respuesta de Arroyo Jiménez en el blog Almacén de Derecho. Vid. Arroyo Jiménez, «Algoritmos y reglamentos». Disponible en: https://almacendederecho.org/algoritmos-y-reglamentos

[393] Boix, «Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones», 257.

extendido en la doctrina para la toma de decisiones automatizada[394]. La cuestión es determinar si esta exigencia respecto del funcionamiento de los modelos algorítmicos es útil, apropiada y justificada en cada contexto normativo[395]. También, y al mismo tiempo, determinar el propio alcance de la transparencia como remedio normativo para abordar los efectos de la opacidad, habitualmente sobreestimado[396].

4.2.1. La transparencia en la propuesta de regulación europea de los sistemas de IA de alto riesgo

La transparencia como "remedio" normativo para la opacidad ha ocupado un lugar central en las propuestas para regular la IA en el ámbito de la UE, así lo expresaba el Libro Blanco de IA: «*La opacidad de los sistemas basados en algoritmos puede abordarse mediante requisitos de transparencia*»[397]. Dicho documento recogió como requisitos de transparencia: la conservación de registros y datos, por un lado, y el suministro o provisión de información, por otro. Mientras que la conservación de registros y datos se enfoca como un requisito de transparencia hacia las autoridades competentes con el fin de verificar de manera efectiva el cumplimiento de las normas aplicables y ejecutarlas[398], el suministro de información tiene por objetivo la promoción del uso responsable de la IA, la creación de confianza y las garantías de reparación cuando proceda, entre los distintos actores del ciclo de vida de la IA[399].

Este modelo es el que sigue, en líneas fundamentales, la propuesta AIA de la Comisión Europea en abril de 2021 para establecer los requisitos obligatorios de los sistemas de IA de alto riesgo. Como tal, se recoge expresamente el requisito de "transparencia y comunicación de información a los usuarios" en el artículo 13, definiendo la forma en la que debe comunicarse la información -concisa, completa, correcta y clara que sea pertinente, accesible y comprensible- en su apartado segundo, y su contenido -instrucciones de uso, características, capacidades y limitaciones de rendimiento del

[394] Lipton, «The Mythos of Model Interpretability»; Zerilli et al., «Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?»

[395] Carabantes, «Black-box artificial intelligence: an epistemological and critical analysis», 8.

[396] Vid. Hert y Lazcoz, «When GDPR-principles blind each other. Accountability, not transparency, at the heart of algorithmic governance».

[397] Comisión Europea, Libro Blanco sobre la inteligencia artificial, 19.

[398] Comisión Europea, 23.

[399] Comisión Europea, 24.

sistema, cambios previsibles y vida útil prevista y medidas de supervisión humana- en el apartado tercero. Este requisito establece la obligación respecto del proveedor desde el diseño y desarrollo del sistema hacia el usuario del mismo, esto es, hacia quien vaya a hacer uso de un sistema de IA bajo su propia autoridad. Se trata, por consiguiente, de un requisito dirigido fundamentalmente a evitar la opacidad procedimental proveedor-usuario[400].

Tal y como argumenta Kiseleva, la propuesta se centra a su vez en la interpretabilidad del sistema -entendida como la posibilidad de comprender y verificar el funcionamiento general del mismo-, sin necesidad de exigir su explicabilidad -entendida como la posibilidad de comprender concretamente las funciones y propiedades que llevan a resultados determinados-[401]. Bajo su punto de vista, el requisito de transparencia no renuncia así a la utilización de modelos con opacidad epistémica o inherente a los mismos. Es decir, la propuesta no limita el uso de los sistemas denominados cajas negras por muy complejos que sean, siempre que el proveedor pueda suministrar la información exigida sobre el funcionamiento y rendimiento del sistema en las condiciones establecidas bajo estas obligaciones de transparencia -y siempre que el sistema de alto riesgo le permita demostrar el cumplimiento normativo en los términos igualmente exigidos por las disposiciones de la propuesta fundamentadas en la rendición de cuentas-.

Por otro lado, la conservación de registros y datos como un requisito de transparencia hacia las autoridades competentes cobra igualmente un gran protagonismo en esta propuesta, aunque no lo haga explícitamente como una "obligación de transparencia"[402].

---

[400] El artículo 52 AIA establece para determinados sistemas de IA obligaciones de suministro de información para los usuarios de los mismos, que deben informar a las personas físicas expuestas a los mismos sobre la interacción o exposición con dichos sistemas. Esto es, se trata de obligaciones de transparencia que impiden el nivel primario de opacidad en el funcionamiento de los sistemas de IA por el cual las personas afectadas conocen la existencia del sistema. Se incluyen aquí los sistemas de IA destinados a interactuar con personas físicas o *chatbots*, sistemas de reconocimiento de emociones o de categorización biométrica y sistemas *deepfakes*, que generen o manipulen contenido de imagen, sonido o vídeo que se asemeje notablemente a personas, objetos, lugares u otras entidades o sucesos existentes, y que puedan inducir erróneamente a una persona a pensar que son auténticos o verídicos.

[401] Kiseleva, «Making AI's transparency transparent: notes on the EU Proposal for the AI Act». Disponible aquí: https://europeanlawblog.eu/2021/07/29/making-ais-transparency-transparent-notes-on-the-eu-proposal-for-the-ai-act/

[402] Este criterio tiene, en realidad, pleno sentido teniendo en cuenta que estos requisitos son parte del sistema de rendición de cuentas o *accountability* diseñado. Los requisitos de transparencia son parte necesaria de este sistema de rendición de cuentas -debiendo definirse qué información se comunica a quién-, si bien, la parte nuclear de este sistema se basa en la relación entre un agente (proveedor en este caso) y un principal (autoridades competentes), en la que el primero debe explicar y justificar su conducta, bajo el juicio del

La documentación técnica tiene por objetivo proporcionar a las autoridades nacionales competentes y los organismos notificados toda la información que necesiten para evaluar si el sistema de IA de alto riesgo cumple con los requisitos obligatorios establecidos para los mismos (art. 11(1) AIA). En el anexo IV de la propuesta se contienen los elementos mínimos que debe contener esta información e incluye, entre otros, información detallada de las medidas y los métodos adoptados para el desarrollo del sistema, especificaciones de diseño como la lógica general y los algoritmos, sobre el sistema de gestión de riesgos, acerca del seguimiento, el funcionamiento y el control del sistema o de los cambios introducidos en su ciclo de vida[403].

En cuanto a las obligaciones de registro, tienen por objeto garantizar un nivel de trazabilidad del funcionamiento del sistema de IA durante su ciclo de vida que resulte adecuado (art. 12(2) AIA). Para ello, se establece el requerimiento de registro automático de determinados eventos por el sistema, que incluye; el periodo de cada uso del mismo, la base de datos de referencia con la que el sistema ha cotejado los datos de entrada, los datos de entrada con los que la búsqueda ha arrojado una correspondencia y la identificación de las personas físicas implicadas en la verificación de los resultados (art. 12 (4) AIA).

Esta documentación y registro obliga, no a comunicar en todo caso toda esta información -al modo de las "obligaciones de transparencia" propiamente dichas-, sino a tenerla disponible por parte del proveedor o permitir que esté disponible para el usuario, de forma que pueda ser comunicada cuando resulte pertinente a las autoridades competentes para la demostración del cumplimiento normativo. De este modo, la transparencia cumple aquí un papel instrumental respecto del sistema normativo de rendición de cuentas diseñado por la Comisión en la propuesta de AIA.

5. Reflexiones provisionales sobre el capítulo primero – tentative thoughts on chapter one

segundo, y soportar las consecuencias de infringir las normas dadas para dicha relación. Vid. sobre el concepto de *accountability* Bovens, «Analysing and Assessing Accountability: A Conceptual Framework».

[403] Para facilitar la actualización de estos elementos y adaptarlos a los posibles avances significativos en el estado del arte, se establece un procedimiento simplificado para la modificación de este anexo en el art. 11(3) AIA.

En este apartado se recogen una serie de reflexiones provisionales a modo de cierre de cada capítulo. Aunque algunas de estas reflexiones servirán de apoyo para las conclusiones de esta investigación, el objetivo de este apartado no es exponer dichas conclusiones propiamente, sino resaltar de forma telegráfica algunos aspectos clave resultado del análisis realizado en cada capítulo.

- La toma de decisiones automatizada basada en la elaboración de perfiles se divide en dos fases fundamentales para su análisis jurídico: (1) el diseño y desarrollo de los modelos algorítmicos y (2) la implementación de los modelos en procesos de toma de decisiones.
- El nivel de automatización de un proceso de decisiones está directamente relacionado con la clase de interacción humana que se establezca en la implementación del modelo, pudiendo responder el tipo de interacción a un mandato del ordenamiento jurídico -intervención humana como mecanismo de gobernanza-.
- El nivel de "automatización" de un proceso de toma de decisiones no es tanto el resultado de un aumento "cuantitativo" de la intervención humana, sino más bien una forma de calificar la forma en que la agencia y el control humanos operan en un contexto en el que se introduce un sistema algorítmico.
- El establecimiento de distintos mecanismos de gobernanza basados en la intervención humana es indispensable para garantizar la supervisión humana como requerimiento central y de obligado cumplimiento en las propuestas europeas para la regulación de la IA de alto riesgo.
- El análisis jurídico parte inevitablemente de una falta de sistematización en la doctrina y jurisprudencia de los distintos mecanismos de gobernanza basados en la intervención humana para la toma de decisiones automatizada y de los distintos objetivos normativos a los que éstos pueden servir.
- El desafío normativo actual consiste en determinar cuál es la participación humana adecuada para las numerosas aplicaciones de los modelos algorítmicos y sus distintos ámbitos de aplicación (¿qué forma debe adoptar la participación?; ¿cuál es el fundamento de esta participación?; ¿para qué sistemas debe ser obligatoria?), y en cómo determinar en el ámbito normativo la cualificación que se exija para dicha participación (¿qué es adecuado, efectivo o significativo?).

- Desde una perspectiva estadística, los sesgos en los procesos de toma de decisiones automatizada no tienen por qué representar un "mal" o un "bien" necesariamente, puesto que suponen una desviación de lo estándar cuyo valor ético o normativo no viene predeterminado.
- Desde una perspectiva ética, los sesgos se han entendido como una distorsión en el análisis que reconfigura la distribución de bienes, servicios, riesgos y oportunidades, o incluso el acceso a la información, de forma moralmente problemática. Es posible que nos encontremos con una redistribución problemática sin necesidad de encontrar una desviación en sentido estadístico.
- Los sesgos han sido hasta el momento un concepto prácticamente desconocido para el Derecho. No obstante, la aparición y reproducción de éstos en la toma de decisiones automatizada basada en la elaboración de perfiles es muy relevante por su capacidad para poner en riesgo bienes jurídicos fundamentales protegidos por el Derecho .
- La forma en la que operan los algoritmos de aprendizaje automático provoca que el riesgo de incurrir en prácticas discriminatorias en el uso de estos sistemas sea considerable. Ahora bien, el análisis jurídico no puede obviar que los sesgos pueden resultar lesivos sin necesidad de afectar a la distribución igualitaria de los bienes protegidos por el Derecho.
- Para poder mitigar los efectos de los sesgos en la toma de decisiones automatizada, y poder así cumplir y demostrar el cumplimiento de los estándares normativos que sean exigibles en cada contexto, es necesario entender las distintas clases de sesgos y su origen e identificar en qué fases de la toma de decisiones se reproducen y de qué manera. La normativa debe, en este sentido, distribuir esta carga entre los distintos actores que participan en el ciclo de vida del sistema.
- El funcionamiento de determinados modelos algorítmicos que se utilizan en la actualidad producen una forma de opacidad característica, denominada aquí “opacidad inherente” y entendida como la incapacidad humana para comprender las correlaciones realizadas por el modelo para alcanzar resultados. Este efecto “caja negra” provoca dificultades, entre otros, para identificar sesgos, para supervisar el funcionamiento de los modelos o para la rendición de cuentas en general.

- Esta opacidad algorítmica ha sido utilizada para encubrir otras formas de opacidad que concurren habitualmente de forma solapada y que pueden producir efectos igualmente nocivos. Resulta indispensable identificar las distintas clases de opacidad para poder aplicar a cada una las soluciones jurídicas que resulten más apropiadas y ponderar, en su caso, los distintos derechos e intereses legítimos que puedan colisionar en la implementación de estos modelos en la toma de decisiones automatizada.
- La transparencia ha sido propuesta como remedio normativo para el efecto caja negra. El análisis jurídico debe abordar con precisión quién revela qué y a quién en esta clase de soluciones, y qué utilidad, justificación y alcance tiene la transparencia para abordar los efectos nocivos de la opacidad en cada contexto de toma de decisiones automatizada.

As a method of recapping each chapter, this section presents a number of tentative thoughts. The research's conclusions will be supported by some of these insights, but the purpose of this section is not to present those conclusions in their entirety. Rather, it aims to emphasize certain key aspects that came out of the analysis done in each chapter in a telegraphic manner.

- For the purposes of legal analysis, automated decision-making based on profiling is divided into two basic stages: (1) the design and development of algorithmic models, and (2) the deployment of the models to decision-making processes.
- The degree to which a decision-making process is automated depends on the kind of human interaction that is established during model implementation, and the kind of interaction may be dictated by a legal requirement (human intervention as a governance mechanism).
- The level of "automation" of a decision-making process is not so much the result of a "quantitative" increase in human intervention, but rather a way of qualifying the way in which human agency and control operate in a context where an algorithmic system is introduced.

- The establishment of different governance mechanisms based on human intervention is indispensable to ensure human oversight as a fundamental and mandatory requirement in European proposals for the regulation of high-risk AI.
- The absence of systematization of the various governance mechanisms based on human intervention for automated decision making and the different regulatory goals they may serve is inexorably the starting point for the legal analysis.
- The current normative challenge is to determine what human involvement is appropriate for the many applications of algorithmic models and their various domains of application (what form should the involvement take; what is the rationale for this involvement; for which systems should it be mandatory?), and how to determine in the normative domain the qualification required for such involvement (what is appropriate, effective or meaningful?).
- From a statistical perspective, biases in automated decision-making processes need not necessarily represent "good" or "evil" since they constitute a variation from the standard whose ethical or normative worth is not predetermined.
- From an ethical perspective, biases have been understood as a distortion in analysis that reconfigures the distribution of goods, services, risks and opportunities, or even access to information, in a morally problematic way. We may encounter problematic redistribution without the need to find bias in a statistical sense.
- Biases have so far been an unknown concept in law. However, the emergence and reproduction of these in automated decision-making based on profiling is highly relevant because of their capacity to jeopardize fundamental legal goods protected by law.
- The way in which machine learning algorithms operate leads to a considerable risk of discriminatory practices in the use of these systems. However, legal analysis cannot ignore the fact that biases can be harmful without affecting the equal distribution of goods protected by law.
- In order to mitigate the effects of biases in automated decision-making, and thus be able to meet and demonstrate compliance with the regulatory standards that may be required in each context, it is necessary to understand the different types of biases and their origin, and to identify at what stages of decision making they

are reproduced and in what way. The regulations must, in this sense, distribute this burden among the different actors involved in the life cycle of the system.

- The operation of certain algorithmic models currently in use produces a characteristic form of opacity, referred to here as “inherent opacity” and understood as the human inability to understand the correlations made by the model to achieve results. This "black box" effect makes it difficult to spot biases, track model performance, or hold different actors accountable in general.
- This algorithmic opacity has been used to cover up other forms of opacity that often occur in an overlapping manner and can produce equally harmful effects. It is necessary to acknowledge the many opacity types in order to apply the best legal solutions to each one and, when necessary, assess the various rights and legitimate interests that can conflict when deploying these models in decision-making processes.
- Transparency has been proposed as a regulatory remedy for the black box effect. Legal analysis must address precisely who discloses what and to whom in these kinds of solutions, and what the usefulness, justification and scope of transparency is in addressing the harmful effects of opacity in each automated decision-making context.

CAPÍTULO 2. TOMA DE DECISIONES AUTOMATIZADA EN EL RGPD: EL ARTÍCULO 22 EN LA UNIDAD DE CUIDADOS INTENSIVOS

CAPÍTULO 2. TOMA DE DECISIONES AUTOMATIZADA EN EL RGPD: EL ARTÍCULO 22 EN LA UNIDAD DE CUIDADOS INTENSIVOS

En el presente capítulo se estudiará la regulación de la toma de decisiones automatizada en el Reglamento General de Protección de Datos (RGPD en adelante), el cual está considerado como la legislación más completa promulgada hasta ahora en esta materia[404]. El ámbito de aplicación de este instrumento normativo es transversal en lo que se refiere al tratamiento de datos personales que se produce en la toma de decisiones automatizada[405], incluida la elaboración de perfiles, y por ello, la doctrina lo ha tomado como referencia para el estudio de los sistemas algorítmicos que realizan esta clase de tratamiento –tanto en derecho público como en derecho privado–[406].

Este capítulo y los posteriores centran su análisis jurídico en la fase de uso del modelo, descartando el análisis de fases anteriores. Por un lado, aunque hemos podido observar que el RGPD es también aplicable al diseño y desarrollo de modelos e introduce aspectos de interés para esta investigación, desde la perspectiva de la toma de decisiones automatizada, su aplicabilidad a estas fases es más bien tangencial o cuanto menos limitada[407]. Por otro lado, debido a que no contamos en este momento con una regulación

[404] Gillis y Simons, «Explanation Justification: GDPR and the Perils of Privacy», 72.

[405] Hay excepciones, por supuesto, como la reconocida en el considerando 19 RGPD para el tratamiento de datos personales conforme a los fines de la Directiva (UE) 2016/680 del Parlamento Europeo y del Consejo, que cuenta también con una regulación específica de la toma de decisiones automatizada y un artículo análogo al 22 RGPD. Vid. Apartado 2.1. Antecedentes y disposiciones análogas en el ámbito normativo europeo. Una predisposición genética evidente en este mismo capítulo.

[406] El RGPD es una de las pocas respuestas legales sólidas que versan sobre las decisiones que toma una máquina sobre un particular, Palma Ortigosa, «Automated Decision-Making in the GDPR. Algorithms in the Scope of the Data Protection». Es obvio que no se trata del único instrumento normativo a través del cual se regula el uso de estas tecnologías, sin embargo, tanto ese carácter transversal como su auxiliaridad en la protección de los derechos fundamentales, le confieren el protagonismo otorgado en la doctrina y en la jurisprudencia disponible en materia de toma de decisiones automatizada. Vid. Janssen, «An approach for a fundamental rights impact assessment to automated decision-making», 82.

[407] Hay quien ha visto en el artículo que obliga a una protección de datos desde el diseño y por defecto -art. 25 RGPD una disposición claramente dirigida a la fase de desarrollo de los sistemas de información: *con el objetivo de garantizar que los intereses relacionados con la privacidad se tengan debidamente en cuenta durante todo el ciclo de vida de dicho desarrollo*. Bygrave, «Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making», 249. Al margen de excepciones poco definidas que abordan la regulación de la tecnología de aprendizaje automático en sus fases de diseño y desarrollo, es innegable que el alcance del RGPD sigue siendo muy limitado en este sentido, no concibiendo estas fases en los modelos algorítmicos como de "producción de conocimiento" sino como de una mera "comunicación de datos", vid. Gellert, «Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms?».

de carácter transversal para el diseño y desarrollo de modelos para la toma de decisiones automatizada[408].

El protagonista ineludible de este análisis es el artículo 22 RGPD que declara el derecho a no ser objeto de decisiones individuales automatizadas, incluida la elaboración de perfiles. Además de abordar directamente el objeto de este estudio, esta disposición contiene los únicos dos mecanismos de gobernanza basados en la intervención humana del RGPD, lo cual pone de manifiesto el interés que contiene para esta investigación.

Como veremos, el ampliamente discutido artículo 22 RGPD contiene severas limitaciones que no pueden obviarse: por un lado, ni éste ni su precedente en la Directiva 95/46/CE han tenido apenas desarrollo jurisprudencial –al menos hasta el momento– y, por otro, la doctrina ha considerado que su alcance es muy limitado. De ahí que sea necesario realizar una lectura conjunta del RGPD y llegar más allá de este artículo 22, analizando otras disposiciones de la norma que introducen otros instrumentos de gobernanza para la toma de decisiones automatizada. Antes de realizar este esfuerzo en capítulos posteriores, es necesario contextualizar sus antecedentes, su ubicación en la normativa de protección de datos y, desde unas líneas más bien generales, su contenido. Este es el objetivo principal del presente capítulo.

El artículo 22 RGPD ha sido caracterizado como un derecho de segunda categoría que permanece inactivo[409], es más, el Reino Unido se estaría planteando prescindir del mismo(!)[410], de ahí la conveniencia de ingresar -en términos metafóricos- a esta disposición en una unidad de cuidados intensivos (UCI).

---

[408] Es previsible que sí contemos con esta regulación próximamente, es por ello, que en este capítulo también se hace referencia a la propuesta de Ley de Inteligencia Artificial y sus precedentes en el seno de la Comisión y el Parlamento europeos.

[409] Bygrave, «Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making», 250.

[410] En el informe «Data: a new direction» del Departament for Digital, Culture Media & Sport se plantea la eliminación está planeando eliminar el derecho a la intervención humana para los sistemas de IA en nombre de la economía y sociedad de datos. Aunque parece descartarse la eliminación completa del artículo 22 RGPD planteada en un informe anterior, esta propuesta pretende ampliar considerablemente el uso de la toma de decisiones automatizada sobre la base de intereses legítimos o intereses públicos. También en relación con los datos personales sensibles. Dado que es probable que el uso de estos sistemas aumente enormemente en muchos sectores, se considera que la necesidad de proporcionar una revisión humana no es viable ni proporcionada. Esta es la principal conclusión que se desprende del análisis del artículo 22 en este informe, que pone de manifiesto su limitada aplicación y la falta de certeza sobre cómo y cuándo se pretenden aplicar en la práctica las salvaguardias actuales. El informe está disponible aquí:

1. La toma de decisiones automatizada en el devenir del derecho a la vida privada y a la protección de datos

En cierto modo es indudable que hemos hecho nuestro el término "privacidad" a la hora de hacer referencia a la protección de nuestra esfera íntima en relación con las tecnologías de la información y comunicación (TIC, en adelante). Son los términos y condiciones sobre nuestra privacidad los que aceptamos al navegar por internet y esa misma privacidad es la que sentimos vulnerada cuando la información personal que albergan nuestros teléfonos móviles escapa de nuestro control, por ejemplo, cuando nos llega publicidad de productos inusualmente específicos o cuando un producto se nos ofrece a distinto precio al consultarlo desde dispositivos distintos.

Sin embargo, nuestro ordenamiento jurídico no sigue un recorrido igualmente intuitivo entre la privacidad, la protección de nuestros derechos fundamentales y el uso de estas tecnologías y nuestros datos personales. El objetivo de este apartado es describir brevemente la relación estrictamente jurídica que mantienen la privacidad, la protección de datos y la intimidad, por supuesto, en relación con el tratamiento masivo de datos para la obtención de inferencias personales que puedan ser fundamento de un proceso de toma de decisiones.

El término privacidad fue acuñado por Warren y Brandeis a finales del siglo XIX[411] y forma parte de la tradición jurídica de los Estados Unidos, sin embargo, y a pesar de su popularidad, "privacidad" no viene recogido o definido por nuestra Constitución, tampoco por el Reglamento General de Protección de Datos (RGPD) o la Ley Orgánica de Protección de Datos y garantía de los derechos digitales (LOPDGDD).

Tampoco es un término utilizado en el marco europeo, a pesar de las múltiples formulaciones que arroja el constitucionalismo comparado[412], aunque puede considerarse que la conjunción de los derechos Carta de Derechos Fundamentales de la Unión Europea (CDFUE) al respeto de la vida privada y familiar -art. 7 CDFUE-, que coincide con la dicción del art. 8.1 del Convenio Europeo de Derechos Humanos (CEDH)) y a la

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1016395/Data_Reform_Consultation_Document__Accessible_.pdf

[411] Warren y Brandeis, «The right to privacy».

[412] Vid. Jove, «Quo vadis, intimidad?», 158.

protección de datos personal (art. 8 CDFUE), vendría a representar el marco de protección que representa la privacidad en Estados Unidos, conformando éstos la base jurídica sobre la que se sustenta la protección de la esfera privada del individuo en el marco de la Unión Europea[413]. No obstante, el valor de la privacidad no se agota en dicha esfera individual, es más, los debates actuales más interesantes centran su interés en la dimensión social y política de la privacidad como elemento esencial de cualquier proyecto democrático que trate de evitar la homogeneización y la exclusión[414].

El apartado cuarto del artículo 18 de la Constitución Española establece que la ley debe limitar el uso de la informática para garantizar el honor y la intimidad personal y familiar de la ciudadanía y el pleno ejercicio de sus derechos[415]. La primera norma que desarrolló este precepto es la derogada Ley Orgánica 5/1992, de 29 de octubre, de regulación del tratamiento automatizado de los datos de carácter personal[416] y en este caso sí, se definió la privacidad en la exposición de motivos, vinculando su concepto a la necesidad de delimitar una nueva frontera de la intimidad y del honor ante la revolución informática que comenzaba a vislumbrarse:

> *El progresivo desarrollo de las técnicas de recolección y almacenamiento de datos y de acceso a los mismos ha expuesto a la privacidad, en efecto, a una amenaza potencial antes desconocida. Nótese que se habla de la privacidad y no de la intimidad: Aquélla es más amplia que ésta, pues en tanto la intimidad protege la esfera en que se desarrollan las facetas más singularmente reservadas de la vida de la persona -el domicilio donde realiza su vida cotidiana, las comunicaciones en las que expresa sus sentimientos, por ejemplo-, la privacidad constituye un conjunto, más amplio, más global, de facetas de su personalidad que, aisladamente consideradas, pueden carecer de significación intrínseca pero que, coherentemente*

[413] Jove, 159.

[414] Turégano Mansilla, «Los valores detrás de la privacidad», 256.

[415] Para Boix, esta limitación del uso de la informática impuesto al legislador para la protección de los derechos y libertades de la ciudadanía expresa, además, una concepción de la evolución tecnológica: como potencialmente muy peligrosa para los derechos de los ciudadanos y pone el acento sobre la necesidad de acompañarlo siempre de las debidas garantías jurídicas. Boix, «Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones», 239.

[416] Hasta su entrada en vigor, las intromisiones ilegítimas derivadas del uso de la informática se regularon por la Ley Orgánica 1/1982, de 5 de mayo, de protección civil del derecho al honor, a la intimidad personal y familiar y a la propia imagen -DT primera-.

enlazadas entre sí, arrojan como precipitado un retrato de la personalidad del individuo que éste tiene derecho a mantener reservado.

La irrupción definitiva de la era digital no hizo sino acentuar la problemática derivada de la reconceptualización del espacio en el que poder estar protegido, de la Mata y Barinas explican que la digitalización provoca que lo público y lo privado confluyan en un espacio social –tanto en su dimensión temporal como espacial– en el cual, «*más que la naturaleza misma del ambiente (virtual o físico) en el que la persona interactúa, juegan un rol trascendental los datos que se exponen y quedan expuestos, la forma en que son tratados y la cuestión de quién tiene acceso a ellos*»[417].

La exposición de motivos arriba citada hizo referencia también al concepto de autodeterminación por el cual toda persona debe tener la posibilidad de determinar el nivel de protección de los datos a ella referentes, con lo que ya en la LO 5/1992, de 29 de octubre, observamos un trazo más o menos reconocible que une los conceptos de privacidad, intimidad y protección de datos; la era informática obliga a la protección de una esfera personal más amplia del espacio tradicionalmente reservado a una intimidad que va incorporando una diversidad de intereses dignos de protección, esto es, la privacidad, y a su vez, el control sobre esa esfera personal se relaciona materialmente con la protección de datos[418]. Ello no implica que la protección de datos esté plenamente integrada en el derecho a la intimidad o vida privada, son derechos autónomos que se solapan parcialmente, habiendo surgido el derecho de protección de datos vinculado al

[417] de la Mata y Barinas Ubiñas, «La privacidad en el diseño y el diseño de la privacidad, también desde el Derecho Penal», 256. Siguiendo esta misma línea argumental, el TJUE reconoce que la injerencia en la privacidad se multiplica debido al papel ubicuo que desempeña Internet en nuestra sociedad -par. 80-. STJUE de 13 de mayo de 2014, asunto C-131/12, Google Spain, S.L. y Google Inc. contra Agencia Española de Protección de Datos (AEPD) y Mario Costeja González.

[418] Asi lo explica Turégano*: Social y jurídicamente la intimidad fue interpretándose esencialmente en su sentido positivo, no ya como aislamiento, cuanto como facultad o poder de revelar y proteger aspectos y circunstancias de nuestra intimidad o nuestras relaciones personales. Además, la intimidad ha ido adquiriendo un sentido más objetivo que subjetivo. Lo privado alude a la información o datos personales que, en cuanto elementos materiales, están más expuestos a la difusión y abuso. El desarrollo acelerado de las tecnologías de la información en las últimas décadas ha puesto el foco en estas facetas de la privacidad otorgándoles cierta autonomía frente a otros aspectos o dimensiones de la misma*. Turégano Mansilla, «Los valores detrás de la privacidad», 264.

derecho a la intimidad[419], también en nuestro ordenamiento[420]. En el artículo 12 de la LO 5/1992, de 29 de octubre, sobre el que volveremos más adelante, se reguló por primera vez la posibilidad de impugnar valoraciones referidas a la personalidad del individuo realizadas exclusivamente sobre el tratamiento automatizado de datos personales.

Con la entrada en vigor en la Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respeta al tratamiento de datos personales y a la libre circulación de estos datos (Directiva 95/46/CE o DPD en adelante) y de su transposición en el ordenamiento jurídico español, la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal (LOPD en adelante), se va consagrando la dimensión social de la privacidad. A partir de dicha dimensión social, el foco sobre el control del individuo y sobre la información que se revela o deja de revelar, se desplaza hacia un control colectivo sobre la justificación de los procesos que recopilan, procesan y usan los datos[421], que se ve reforzado con un incipiente derecho a la protección de datos que refuerza, a su vez, este control sobre los principios de transparencia y responsabilidad en el tratamiento de los datos personales, con independencia de que éstos pertenezcan o no a la esfera privada de la persona interesada[422].

Entre las cuestiones que afectan a esa recopilación, procesamiento y uso de los datos, la DPD regula por primera vez la toma de decisiones automatizadas en el marco normativo europeo de la protección de datos en su artículo 15. Solove utilizó como ejemplo esta

[419] van Bekkum y Borgesius, «Digital welfare fraud detection and the Dutch SyRI judgment», 14. Autonomía recogida de forma expresa en los artículos 7 y 8 de la Carta de los Derechos Fundamentales de la UE. Al respecto, más creativo ha tenido que ser el TEDH estirando el derecho a la vida privada y familiar para incluir la protección de datos, dado que el CEDH data de 1950 y no incluye un derecho de protección de datos, vid. Gutwirth y De Hert, «Regulating Profiling in a Democratic Constitutional State», 258 y ss.

[420] Así nuestro Tribunal Constitucional, en sus sentencias 290/2000 y 292/2000, estableció con claridad -FJ. 5-: *Este derecho fundamental a la protección de datos, a diferencia del derecho a la intimidad del art. 18.1 CE, con el cual comparte el objetivo de ofrecer una protección constitucional eficaz de la vida privada personal y familiar, atribuye al titular una serie de facultades que consiste, en la mayor parte, en el poder jurídico de imponer a terceros la realización o la omisión de determinados comportamientos la regulación concreta de los cuales tiene que establecer la Ley, aquella que, de acuerdo con el art. 18.4 CE, tiene que limitar el uso de la informática, bien desarrollando el derecho fundamental a la protección de datos (art. 81.1 CE), bien regulando el ejercicio (art. 53.1 CE). La peculiaridad de este derecho fundamental a la protección de datos respecto de aquel derecho fundamental tan afín cómo es el de la intimidad radica, así pues, en la distinta función que hacen, cosa que implica, por consiguiente, que también el objeto y el contenido difieran*. STC 292/2000, de 30 de noviembre de 2000, núm. recurso 1.463/2000.

[421] Turégano Mansilla, «Los valores detrás de la privacidad», 276.

[422] Lynskey, «Deconstructing data protection: The "added-value" of a right to data protection in the eu legal order».

disposición para señalar que la Unión Europea a través de la misma reconocía ya las dimensiones de la problemática que denominó *database privacy problem*, problemática que la normativa estadounidense obviaba por completo[423]. Dicha problemática consistía en que las bases de datos y su tratamiento alteran la forma en que se toman decisiones y elaboran juicios que afectan a nuestras vidas, exacerbando y transformando los desequilibrios de poder existentes en nuestras relaciones con las instituciones burocráticas[424]. En definitiva, esa capacidad para decidir y controlar las informaciones personales que sobre nosotros conocen los demás y su utilización, se veía mermada por esta nueva problemática y la DPD pretendía devolver el control social e individual de estos usos, entre otros, a través de la regulación de la toma de decisiones automatizada.

Según Solove, esta problemática se comprende mejor a partir del relato de *El Proceso* de Franz Kafka, que en la manida metáfora del gran hermano que aparece en *1984* de George Orwell[425]. *El Proceso* narra la pesadilla vivida por su protagonista, Josef K., a partir del día en el que se le informa inesperadamente de que está detenido sin darle ningún tipo de razón. Un organismo burocrático ha abierto un expediente frente a él, pero Josef K. no tiene acceso a la información o los motivos tras esta incidencia. A lo largo del resto de la novela, el protagonista intenta desesperadamente averiguar por qué este organismo se interesa por su vida, pero su búsqueda es inútil ante una burocracia clandestina y laberíntica, en la cual ni siquiera los funcionarios con los que tiene contacto saben explicarle el porqué de su situación.

Solove acude a esta metáfora desde una amplia perspectiva que pretende ahondar en el concepto de privacidad o, mejor dicho, reformular el concepto de privacidad dada la disrupción que ha introducido la utilización de bases de datos y su tratamiento por parte de poderes burocráticos -grandes organizaciones públicas y privadas con estructuras jerárquicas y cuyo funcionamiento se organiza de acuerdo a un conjunto de reglas, rutinas y procedimientos previamente establecidos- para la evaluación de aspectos personales y

[423] Solove, «Privacy and Power: Computer Databases and Metaphors for Information Privacy», 1460.

[424] Solove, 1399.

[425] Solove, 1398. También siguiendo con metáforas que inciden sobre la vigilancia permanente y masiva (e invisible en la mayoría de ocasiones) se utiliza habitualmente el panóptico de Bentham. Vid. Garriga Domínguez, «La elaboración de perfiles y su impacto en los derechos fundamentales. Una primera aproximación a su regulación en el reglamento general de protección de datos de la Unión Europea», 122 y ss.

la toma de decisiones sobre las mismas[426]. El uso de esta metáfora provocó que el artículo 15 DPD fuese conocido como la disposición kafkiana[427].

A pesar del fracaso de esta disposición, fracaso sobre el que entraremos más adelante, el RGPD recogió nuevamente una disposición similar en su artículo 22. Sin embargo, toda esta problemática iba a adquirir una dimensión ulterior a lo largo de la pasada década que la nueva normativa europea de protección de datos debía estar preparada para abordar[428]. La introducción de complejos algoritmos de aprendizaje automático para el tratamiento de ingentes bases de datos con fines de perfilado individual y como base para la toma de decisiones de organizaciones públicas y privadas, supuso un nuevo hito en la privacidad y la protección de datos personales[429].

El desequilibrio de poder existente en el *database privacy problem* definido por Solove, se ve acentuado como consecuencia de diversos factores, no todos de carácter técnico,

[426] Solove, *The Digital Person: Technology and Privacy in the Information Age*, 27 y ss.

[427] Más adelante, a través de esta metáfora, Korff se refiere ya a los riesgos que introduce la elaboración de perfiles -término al que Solove no hace referencia explícita y que tampoco aparece como tal regulada por la DPD- sobre consumidores y ciudadanas por poderosas corporaciones y organismos estatales –nuevamente incide en la importancia del desequilibrio en las relaciones de poder–, perfiles sobre los que se toman decisiones que les afectan de forma significativa, sin que los responsables sepan explicar el razonamiento subyacente a las mismas y denegando a dichas personas cualquier recurso individual o colectivo efectivo. Korff, «Comments on Selected Topics in the Draft EU Data Protection Regulation», 25; Vid. también Hildebrandt, «Technology and the End of Law». En los análisis posteriores ya referidos al RGPD vuelve a utilizarse esta analogía, Selbst y Powles relacionando el relato kafkiano con el derecho a recibir información significativa en el marco de la toma de decisiones automatizada del Reglamento; dichos autores entienden que las explicaciones pueden tener tanto un valor instrumental, como un valor intrínseco de la autonomía, y este último obedece a la necesidad de una persona de tener libre albedrío y control, anhelo de Josef K. reflejado a lo largo de la novela, vid. Selbst y Powles, «Meaningful information and the right to explanation», 236; también en Zuiderveen Borgesius, «Strengthening legal protection against discrimination by algorithms and artificial intelligence».

[428] Supervisor Europeo de Protección de Datos (SEPD), «Dictamen 3/2015 -La gran oportunidad de Europa. Recomendaciones del SEPD sobre las opciones de la UE en cuanto a la reforma de la protección de datos-». Así lo expone: *A juzgar por la longevidad de la Directiva 95/46/CE, es razonable prever un plazo similar antes de que se produzca la siguiente revisión importante de la normativa de protección de datos, acaso no antes del decenio de 2030. Mucho antes de dicha fecha, cabe prever una convergencia de las tecnologías basadas en los datos con la inteligencia artificial, el procesamiento de lenguajes naturales y los sistemas biométricos, lo que permitirá el desarrollo de aplicaciones dotadas de una capacidad de aprendizaje artificial y una inteligencia avanzada. Estas tecnologías suponen un desafío para los principios de la protección de datos. Una reforma orientada hacia el futuro debe fundamentarse, pues, en la dignidad de la persona y estar basada en criterios éticos*.

[429] Para Garriga el perfil informático en el contexto de ubicuidad de las tecnologías de la información instaura un determinismo incompatible con la autodeterminación del individuo que pasa a ser un mero objeto de información, dejando de ser un dotado de dignidad y sujeto de derechos fundamentales. Garriga Domínguez, «La elaboración de perfiles y su impacto en los derechos fundamentales. Una primera aproximación a su regulación en el reglamento general de protección de datos de la Unión Europea», 138.

que ya se han identificado anteriormente[430], lo cual significaba que la regulación de la toma de decisiones automatizada de esta disposición sería crucial para la nueva regulación[431]. El Supervisor Europeo de Datos Personales (SEPD, en adelante) hace referencia a que la reutilización algorítmica hace que los datos pierdan su contexto original, afectando a la autodeterminación informativa de la persona y reduciendo aun más el control de las interesadas sobre sus datos[432].

En efecto, el RGPD hace frente a unos retos diferentes de los enfrentados por la DPD, además, lo hace ya sobre la base exclusiva del derecho fundamental a la protección de datos consagrado en el artículo 8, apartado 1, de la CDFUE y artículo 16, apartado 1, del Tratado de Funcionamiento de la Unión Europea (TFUE)[433] y supone un avance considerable y una toma de conciencia de las amenazas que para los derechos fundamentales provocan estas tecnologías[434].

---

[430] Escala y velocidad del tratamiento de datos, distintos niveles opacidad y sesgos algorítmicos, ubicuidad de los sistemas de vigilancia y toma de decisiones, impacto sobre grupos *ad hoc* que trascienden la concepción tradicional de grupo, etc. Bygrave hacía ya referencia en el año 2001 al conjunto de causas socioeconómicas y técnicas que provocaban esta transformación, no solo la frecuencia, la intensidad y el alcance de las prácticas de perfilado por parte de grandes organizaciones burocráticas estaban en constante crecimiento, convirtiendo la elaboración de perfiles en una industria emergente por derecho propio, sino que las técnicas en las que se basaba eran también cada vez más sofisticadas. Vid. Bygrave, «Minding the machine: art 15 of the EC Data Protection Directive and automated profiling».

[431] Korff, «Comments on Selected Topics in the Draft EU Data Protection Regulation», 19.

[432] Supervisor Europeo de Protección de Datos (SEPD), «Dictamen 3/2015 -La gran oportunidad de Europa. Recomendaciones del SEPD sobre las opciones de la UE en cuanto a la reforma de la protección de datos-», 15.

[433] Rallo Lombarte, «El nuevo derecho de protección de datos», 49. El Considerando 4 RGPD mantiene la única mención al derecho a la vida privada y familiar en este contexto normativo: *El tratamiento de datos personales debe estar concebido para servir a la humanidad. El derecho a la protección de los datos personales no es un derecho absoluto sino que debe considerarse en relación con su función en la sociedad y mantener el equilibrio con otros derechos fundamentales, con arreglo al principio de proporcionalidad. El presente Reglamento respeta todos los derechos fundamentales y observa las libertades y los principios reconocidos en la Carta conforme se consagran en los Tratados, en particular el respeto de la vida privada y familiar, del domicilio y de las comunicaciones, (…)*. En lo que respecta a nuestro ordenamiento jurídico, dice Rallo Lombarte: *La convivencia del art. 8 CDFUE y del art. 18.4 CE está pacíficamente garantizada por vía hermenéutica en la medida en que el derecho fundamental garantizado por el art. 18.4 CE va a ser directa y principalmente regulado por el RGPD desplazándose el canon de protección del derecho fundamental a la interpretación que del art. 8 CDFUE haga el TJUE*. Rallo Lombarte, 58.

[434] Sancho Lopez, «Legal Strategies To Ensure Fundamental Rights in Front of the Chanllenges of Big Data», 5. A la luz de estas amenazas, tampoco puede desmerecerse el papel "constitucional" que el TEDH ha adoptado en los últimos años en su interpretación del derecho al respeto a la vida privada y familiar: *el Tribunal, al menos en este tipo de casos, se ha transformado de un tribunal tradicional de derechos humanos, que evalúa en concreto si se ha violado uno o varios de los derechos humanos de un solicitante y, en caso afirmativo, si se debe conceder una indemnización por daños y perjuicios, a un tribunal constitucional, que evalúa las leyes en abstracto y las pone a prueba en función de los principios generales derivados del Estado de Derecho y la separación de poderes*. Vid. van der Sloot, «The Quality of Law:

Ahora bien, el RGPD, no es un mero instrumento normativo para el reconocimiento y ejercicio de derechos que permite el control individual de los datos personales, sino que representa el control colectivo sobre la justificación de los procesos que recopilan, procesan y usan los datos. Y ello se traduce especialmente en el establecimiento de una sólida responsabilidad sobre el cumplimiento normativo y su demostración como piedra angular del RGPD, también en la regulación de la toma de decisiones automatizada, como veremos.

Parte de la doctrina, especialmente desde fuera de Europa, ha sido muy crítica con la inclusión del artículo 22 RGPD ante este desarrollo tecnológico. Para Zarsky, estamos ante el ejemplo más ilustrativo por el cual el RGPD rechaza la revolución del *Big Data* y su innovación, alega que reclamar interpretabilidad pone en peligro la precisión de los modelos y que la intervención humana entorpece y ralentiza los procesos automatizados de las tecnologías de vanguardia[435]. También el artículo 22 ha sido catalogado como un elemento del RGPD que podría interpretarse como un derecho contra el tratamiento automatizado[436].

Aunque, del otro lado, se ha argumentado que las incontables excepciones que contiene esta disposición reflejan cómo los EEMM no quisieron renunciar a su *hambre* por los datos masivos y el perfilado algorítmico, especialmente si éste se produce en un contexto de innovación, a pesar de su potencial impacto para la ciudadanía[437]. Profundicemos en el análisis de esta disposición para valorar con mayor criterio estas posiciones.

2. La disposición kafkiana: Artículo 22 del RGPD. Razones para su ingreso en la UCI.

Este apartado se centra exclusivamente en la disposición 22 del RGPD más que en el conjunto de disposiciones que regula la toma de decisiones en el mismo. Las razones que obligan a ello son, principalmente, la multitud de ambigüedades que aún hoy persisten en torno a su aplicación y que merecen al menos tenerse en cuenta antes de abordar un

How the European Court of Human Rights gradually became a European Constitutional Court for privacy cases».

[435] Zarsky, «Incompatible: The GDPR in the Age of Big Data», 1017-18.

[436] Huq, «A Right to a Human Decision», 623.

[437] Hilden, «The Politics of Datafication: The influence of lobbyists on the EU's data protection reform and its consequences for the legitimacy of the General Data Protection Regulation», 197-98.

análisis más amplio de la regulación de la toma de decisiones automatizada en el RGPD. Las razones que han llevado a esta disposición a este punto crítico no pueden obviar los antecedentes de la misma, así como otras disposiciones análogas que encontramos en vigor en otros ámbitos normativos. Por ello, el análisis del fracaso del artículo 22 RGPD debe comenzar, necesariamente, por su predisposición histórica -léase genética- al fracaso.

1.1. Antecedentes y disposiciones análogas en el ámbito normativo europeo. Una predisposición genética evidente.

En el ámbito europeo, antes del mencionado artículo 15 DPD, la ley francesa relativa a la informática, ficheros y libertades[438] incorporó en primer lugar garantías frente a la elaboración de perfiles y las decisiones automatizadas en 1978, tanto para decisiones judiciales, administrativas o del ámbito privado[439], bajo el fundamento de la protección de la dignidad humana bajo el determinismo tecnológico[440]. Siguiendo esta normativa, nuestro ordenamiento fue uno de los pocos, junto a Portugal[441], que también reguló esta figura en el ámbito europeo anticipándose al derecho de la Unión. El artículo 12 de la LO 5/1992, de 29 de octubre, albergó un derecho a impugnar valoraciones basadas exclusivamente en datos automatizados:

> *El afectado podrá impugnar los actos administrativos o decisiones privadas que impliquen una valoración de su comportamiento cuyo único fundamento sea un tratamiento automatizado de datos de carácter personal que ofrezca una definición de sus características o personalidad.*

[438] Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

[439] El artículo 2 de la Loi n° 78-17 en su primera versión: *Ninguna decisión judicial que implique una evaluación del comportamiento humano puede basarse en el tratamiento automatizado de información que ofrezca un perfil o una definición de la personalidad del interesado. Ninguna decisión administrativa o privada que implique una evaluación del comportamiento humano puede basarse únicamente en el tratamiento automatizado de información que ofrezca un perfil o una definición de la personalidad del interesado*. Su artículo 3 recogía un derecho de información e impugnación: *Toda persona tiene derecho a conocer y a impugnar la información y el razonamiento utilizados en las operaciones de tratamiento automatizado cuyos resultados se utilizan en su contra*. Traducción del original en francés disponible aquí: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460/1979-09-22/

[440] Bygrave, «Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making», 249.

[441] Artículo 16 de la derogada Lei no 10/91 de 12. de Abril 1991, da Proteccão de Dados Pessoais face à Informática.

Estos antecedentes dieron lugar al precedente del actual artículo 22 RGPD, esto es, el artículo 15 de la ya derogada Directiva 95/46/CE, o la genuina disposición kafkiana. Lamentablemente, contenía muchos límites y varias condiciones no exentas de ambigüedad, y no pasó de ser considerado un derecho de segundo rango, sin que el Tribunal de Justicia de la UE (TJUE en adelante) o las jurisdicciones nacionales resolviesen casos relevantes de aplicación del mismo[442]. Tampoco en las resoluciones de las autoridades de control de los EEMM[443]. El contenido de la disposición era el siguiente[444]:

Artículo 15

Decisiones individuales automatizadas

1. Los Estados miembros reconocerán a las personas el derecho a no verse sometidas a una decisión con efectos jurídicos sobre ellas o que les afecte de manera significativa, que se base únicamente en un tratamiento automatizado de datos destinado a evaluar determinados aspectos de su personalidad, como su rendimiento laboral, crédito, fiabilidad, conducta, etc.

2. Los Estados miembros permitirán, sin perjuicio de lo dispuesto en los demás artículos de la presente Directiva, que una persona pueda verse sometida a una de las decisiones contempladas en el apartado 1 cuando dicha decisión:

---

[442] Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 30. Exceptuando la sentencia de 2014, de la que se hará mención más adelante, del Tribunal Federal de Justicia de Alemania o Bundesgerichtshof (BGH) para el caso SCHUFA sobre valoraciones crediticias en aplicación del artículo 15 DPD. Vid. apartado 1.2. Intervención humana significativa: la necesidad de superar un concepto formal de intervención humana, en el capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica

[443] Bygrave, «Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making», 250. También como excepción, Bygrave hace referencia a la Décision 2017-053 du 30 août 2017 de la Commission nationale de l'informatique et des libertés (CNIL), que resolvió que un sistema automatizado para determinar la admisión en las universidades francesas infringía la normativa que transpuso el artículo 15 de la DPD.

[444] La transposición de esta norma en el ordenamiento jurídico español se recogió en el artículo 13 de la también derogada Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal, conservando en el título el derecho a la impugnación de valoraciones al que hacía referencia el artículo 12 de la LO 5/1992, de 29 de octubre: *Artículo 13. Impugnación de valoraciones. 1. Los ciudadanos tienen derecho a no verse sometidos a una decisión con efectos jurídicos, sobre ellos o que les afecte de manera significativa, que se base únicamente en un tratamiento de datos destinados a evaluar determinados aspectos de su personalidad.; 2. El afectado podrá impugnar los actos administrativos o decisiones privadas que impliquen una valoración de su comportamiento, cuyo único fundamento sea un tratamiento de datos de carácter personal que ofrezca una definición de sus características o personalidad.; 3. En este caso, el afectado tendrá derecho a obtener información del responsable del fichero sobre los criterios de valoración y el programa utilizados en el tratamiento que sirvió para adoptar la decisión en que consistió el acto.; 4. La valoración sobre el comportamiento de los ciudadanos, basada en un tratamiento de datos, únicamente podrá tener valor probatorio a petición del afectado.*

a) se haya adoptado en el marco de la celebración o ejecución de un contrato, siempre que la petición de celebración o ejecución del contrato presentada por el interesado se haya satisfecho o que existan medidas apropiadas, como la posibilidad de defender su punto de vista, para la salvaguardia de su interés legítimo; o

b) esté autorizada por una ley que establezca medidas que garanticen el interés legítimo del interesado. -nota-

Esta disposición introdujo una base jurídica sobre la que establecer una serie de estrategias para mitigar los riesgos de la elaboración de perfiles automatizados que aumentaban, con una serie de mecanismos de gobernanza ex ante y ex post, las obligaciones de transparencia y responsabilidad del responsable del tratamiento, sin embargo, esta base jurídica nunca ocupó el lugar destacado que merecía en el debate jurídico[445].

Su estructura estaba basada en la proclamación de un derecho -15(1) DPD- que admite excepciones bajo determinados supuestos -15(2) DPD-. Cuando el responsable del tratamiento se amparaba en alguna de estas excepciones para adoptar decisiones basadas únicamente en el tratamiento automatizado, se le obligaba a adoptar una serie de medidas de salvaguarda y es, probablemente, lo que se ha mantenido de forma más fiel en el actual artículo 22 RGPD. Quizás reproducir esta estructura sea el primer error cometido por el legislador europeo, al igual que yerra al reproducir algunos aspectos cuya ambigüedad era ya cuestionada con anterioridad: ¿se trata de un *derecho-facultad* del interesado o de una *prohibición general*? ¿qué son los *efectos* jurídicos o de afectación similar? ¿qué implica que la decisión esté basada *únicamente* en el tratamiento automatizado? Sobre todas estas cuestiones volveremos más adelante.

Además de los antecedentes arriba descritos, encontramos en la regulación vigente varias disposiciones análogas que merecen nuestra atención[446]. En el ámbito de la UE, por un lado, tenemos la Directiva (UE) 2016/680 relativa al tratamiento de datos personales por parte de las autoridades para fines de prevención, investigación, detección o

[445] Schermer, «The limits of privacy in automated profiling and data mining», 52.

[446] No se incluyen aquí disposiciones que hayan sido aprobadas por los Estados Miembro en desarrollo de las excepciones contenidas en el propio artículo 22 (22(2)(b) y 22(4)), como análisis de derecho comparado en esta materia se recomienda Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations».

enjuiciamiento penales, que contiene en su artículo 11 el derecho a no ser objeto de una decisión que evalúe aspectos personales que le conciernen que se base únicamente en un tratamiento automatizado de los datos[447]. Asimismo, la Directiva (UE) 2016/681 relativa a la utilización de datos del registro de nombres de los pasajeros, incluye en su artículo 7(6) la prohibición de tomar ninguna decisión que pudiera tener efectos jurídicos adversos para una persona o afectarle gravemente en razón únicamente del tratamiento automatizado de datos del registro[448].

En el ámbito del Consejo de Europa, el Convenio 108[449] fue un instrumento pionero para la regulación de datos a nivel internacional, siendo durante más de 15 años -hasta la entrada en vigor de la DPD- la única normativa supranacional vinculante en materia de protección de datos[450]. La versión original de 1981 no ofrecía disposiciones análogas que regulasen la toma de decisiones automatizada, sin embargo, en el protocolo de enmienda que dio lugar al conocido como Convenio 108+[451], su artículo 9 recoge el derecho de toda

[447] Artículo 11 de la Directiva (UE) 2016/680. Mecanismo de decisión individual automatizado: *1. Los Estados miembros dispondrán la prohibición de las decisiones basadas únicamente en un tratamiento automatizado, incluida la elaboración de perfiles, que produzcan efectos jurídicos negativos para el interesado o le afecten significativamente, salvo que estén autorizadas por el Derecho de la Unión o del Estado miembro a la que esté sujeto el responsable del tratamiento y que establezca medidas adecuadas para salvaguardar los derechos y libertades del interesado, al menos el derecho a obtener la intervención humana por parte del responsable del tratamiento.; 2. Las decisiones a que se refiere el apartado 1 del presente artículo no se basarán en las categorías especiales de datos personales contempladas en el artículo 10, salvo que se hayan tomado las medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado.; 3. La elaboración de perfiles que dé lugar a una discriminación de las personas físicas basándose en las categorías especiales de datos personales establecidas en el artículo 10 quedará prohibida, de conformidad con el Derecho de la Unión.* El contenido de dicho artículo 11 ha sido objeto de transposición de forma prácticamente literal por la Ley Orgánica 7/2021, de 26 de mayo, de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales. Para un análisis sobre este artículo, vid. Guzman Fluja, «Proceso penal y justicia automatizada».

[448] Artículo 7 de la Directiva 2016/681. Autoridades competentes: *6. Las autoridades competentes no adoptarán ninguna decisión que produzca efectos jurídicos adversos para una persona o que afecte significativamente a una persona únicamente en razón del tratamiento automatizado de datos PNR. Dichas decisiones no deberán basarse en la raza o el origen étnico, las opiniones políticas, las creencias religiosas o filosóficas, la pertenencia a un sindicato, la salud o la vida u orientación sexual de la persona.* En este caso, la disposición también fue objeto de transposición literal en su artículo 14.6 por la Ley Orgánica 1/2020, de 16 de septiembre, sobre la utilización de los datos del Registro de Nombres de Pasajeros para la prevención, detección, investigación y enjuiciamiento de delitos de terrorismo y delitos graves.

[449] Convenio para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal, hecho en Estrasburgo el 28 de enero de 1981 (BOE-A-1985-23447).

[450] de Hert y Papakonstantinou, «Framing Big Data in the Council of Europe and the EU data protection law systems: Adding 'should' to 'must' via soft law to address more than only individual harms», 4.

[451] Protocolo de enmienda del Convenio del Consejo de Europa para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal, hecho en Estrasburgo el 10 de octubre de 2018 (Convenio número 223 del Consejo de Europa).

persona a no ser objeto de una decisión que la afecte significativamente sobre la base exclusiva de un tratamiento automatizado de datos sin que se tengan en cuenta sus opiniones[452].

ROIG sostiene que este artículo sigue claramente las previsiones del artículo 22 RGPD[453], aunque la doctrina ha señalado, por ejemplo, que amplía el ámbito de aplicación de los derechos de información reconociendo de forma explícita un derecho a una explicación y no solo para las decisiones basadas únicamente en el tratamiento automatizado[454]. En cualquier caso, estamos ante una disposición análoga que también se aplica en muchos EEMM -cada vez más según ratifiquen el protocolo de enmienda Convenio 108+- y que por tanto será también relevante en el ámbito de la protección de datos.

3. Análisis de su contenido. El diagnóstico a debate.

A continuación, se desarrolla un sucinto análisis del contenido del artículo 22 RGPD. El análisis se fundamenta en las siguientes fuentes: las Directrices sobre decisiones individuales automatizadas y elaboración de perfiles del Grupo de Trabajo del Artículo 29 (GT29 en adelante)[455]; la jurisprudencia de la que disponemos en el TJUE y en los tribunales de los diferentes EEMM hasta el momento[456]; la doctrina europea accesible en

[452] Artículo 9 del Convenio 108+: *1. Toda persona física tendrá derecho a: a. no ser objeto de una decisión que la afecte significativamente sobre la base exclusiva de un tratamiento automatizado de datos sin que se tengan en cuenta sus opiniones; (…) c. recibir una explicación, cuando así lo solicite, sobre el motivo subyacente al tratamiento de datos cuando se le apliquen los resultados de dicho tratamiento; d. oponerse en cualquier momento, por motivos relacionados con su situación, al tratamiento de los datos de carácter personal que la conciernan, a menos que el responsable del tratamiento demuestre que existen motivos legítimos para dicho tratamiento que prevalezcan sobre los intereses o derechos y libertades fundamentales del afectado; (…) 2. La letra a del párrafo 1 no se aplicará cuando la decisión esté autorizada por el ordenamiento al que esté sujeto el responsable del tratamiento y dicho ordenamiento establezca también medidas adecuadas para salvaguardar los derechos, las libertades y los intereses legítimos de la persona concernida*. Texto extraído de: https://www.congreso.es/public_oficiales/L14/CORT/BOCG/A/BOCG-14-CG-A-34.PDF

[453] Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 33.

[454] de Hert y Papakonstantinou, «Framing Big Data in the Council of Europe and the EU data protection law systems: Adding 'should' to 'must' via soft law to address more than only individual harms», 7.

[455] Dichas directrices fueron refrendadas por el Comité Europeo de Protección de Datos (CEPD en adelante) en su primera sesión plenaria: Endorsement 1/2018, Bruselas, 25 de mayo de 2018. Disponible aquí: https://edpb.europa.eu/news/news/2018/endorsement-gdpr-wp29-guidelines-edpb_es

[456] Para recopilar la jurisprudencia disponible he acudido al repositorio GDPRhub. Disponible aquí: https://gdprhub.eu/index.php?title=Welcome_to_GDPRhub. De especial relevancia son los casos resueltos por el Tribunal de Distrito de La Haya -SyRI- y por el Tribunal de Distrito de Ámsterdam -Uber y Ola- que he tenido ocasión de analizar en mayor profundidad en trabajos previos.

lengua inglesa y castellana[457]; así como la jurisprudencia y doctrina más relevante acerca de los antecedentes y disposiciones análogas desarrolladas en el apartado anterior.

Este análisis se centra exclusivamente en las dos cuestiones más ambiguas, así como en las principales limitaciones de la disposición que tienen peso sobre la interpretación que se propone más adelante. Otras cuestiones, que también pueden resultar de interés, simplemente se citan y se hace referencia a otros trabajos sobre las mismas, más extensos y mejor documentados. Tras la exposición de la ubicación y sinuosa estructura de este artículo, los siguientes apartados tratan de responder a los interrogantes: ¿recoge este derecho una prohibición general o un derecho a interponer por la persona interesada?, y ¿qué clase de efectos han de concurrir para que sea aplicable este derecho?

3.1. Ubicación de la toma de decisiones automatizada y de la elaboración de perfiles en el RGPD

En este apartado, resulta oportuno echar un vistazo a la disposición y a su ubicación en el RGPD. No se realizará una contextualización más exhaustiva del Reglamento como norma dado que excede el ámbito de esta investigación[458].

El capítulo III del Reglamento recoge los 'Derechos del interesado' entre los cuales encontramos obligaciones para el responsable del tratamiento respecto de las personas interesadas que habitualmente requieren de una petición o solicitud por las mismas, si bien también hay obligaciones que el responsable debe llevar a cabo con carácter *ex officio*, sin necesidad de que la persona interesada realice petición alguna. Entre los distintos derechos recogidos en este capítulo, divididos por secciones, tenemos: los derechos de información y acceso (arts. 13-15), los derechos de rectificación y supresión, entre los que se incluyen además el derecho a la limitación del tratamiento y el derecho de portabilidad de los datos (arts. 16-20) y, por último, el derecho de oposición y decisiones individuales automatizadas (arts. 21-22). Además, se incluyen en este capítulo las condiciones a las que el responsable debe ajustarse en virtud del principio de

[457] No he seguido una metodología concreta para la búsqueda de esta bibliografía que resulte reseñable.

[458] Para ello contamos, además, con monografías de excelente calidad. Cuando sea preciso, a lo largo del presente capítulo, se realizarán otras aproximaciones que se consideren necesarias para profundizar en aspectos de carácter específico. Se recomienda, por su extensión y calidad, la obra dirigida por Troncoso Reigada (Dir.), *Comentario al Reglamento General de Protección de Datos y a la Ley Orgánica de Protección de Datos personales y Garantía de los Derechos Digitales*.

transparencia para el ejercicio de los derechos enumerados anteriormente (art. 12), así como las limitaciones que por el Derecho de la Unión o de los EEMM podrían llegar a establecerse para el ejercicio de los mismos (art. 23)[459].

Acerca del artículo 22 versa un extenso Considerando 71 RGPD que aborda, a partir de algunos ejemplos el derecho a no ser objeto a una decisión automatizada, las excepciones al mismo y bajo qué condiciones pueden darse, el impacto de esta clase de decisiones y, en particular, de la elaboración de perfiles, y las obligaciones del responsable del tratamiento al realizar esta clase de tratamiento conforme a los principios del RGPD. Dice así:

> *El interesado debe tener derecho a no ser objeto de una decisión, que puede incluir una medida, que evalúe aspectos personales relativos a él, y que se base únicamente en el tratamiento automatizado y produzca efectos jurídicos en él o le afecte significativamente de modo similar, como 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. Este tipo de tratamiento incluye la elaboración de perfiles consistente en cualquier forma de tratamiento de los datos personales que evalúe aspectos personales relativos a una persona física, en particular para analizar o predecir aspectos relacionados con el rendimiento en el trabajo, la situación económica, la salud, las preferencias o intereses personales, la fiabilidad o el comportamiento, la situación o los movimientos del interesado, en la medida en que produzca efectos jurídicos en él o le afecte significativamente de modo similar. Sin embargo, se deben permitir las decisiones basadas en tal tratamiento, incluida la elaboración de perfiles, si lo autoriza expresamente el Derecho de la Unión o de los Estados miembros aplicable al responsable del tratamiento, incluso con fines de control y prevención del fraude y la evasión fiscal, realizada de conformidad con las reglamentaciones, normas y recomendaciones de las instituciones de la Unión o de los órganos de supervisión nacionales y para garantizar la seguridad y la fiabilidad de un servicio prestado por el responsable del tratamiento, o necesario para la conclusión o ejecución de un contrato entre el interesado y un responsable del tratamiento, o en los casos en los que el interesado haya dado su consentimiento*

[459] Cuyo espíritu básicamente replica las posibilidades de limitación del derecho al respeto de la vida privada y familiar recogidas en el apartado 2 del artículo 8 del CEDH, ampliamente desarrolladas en la jurisprudencia del TEDH. Muy recomendable al respecto la Guía del TEDH sobre dicho artículo actualizada a 31 de diciembre de 2020. Disponible aquí: https://www.echr.coe.int/documents/guide_art_8_eng.pdf

explícito. En cualquier caso, dicho tratamiento debe estar sujeto a las garantías apropiadas, entre las que se deben incluir la información específica al interesado y el 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. Tal medida no debe afectar a un menor.

A fin de garantizar un tratamiento leal y transparente respecto del interesado, teniendo en cuenta las circunstancias y contexto específicos en los que se tratan los datos personales, el responsable del tratamiento debe utilizar procedimientos matemáticos o estadísticos adecuados para la elaboración de perfiles, aplicar medidas técnicas y organizativas apropiadas para garantizar, en particular, que se corrigen 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. Las decisiones automatizadas y la elaboración de perfiles sobre la base de categorías particulares de datos personales únicamente deben permitirse en condiciones específicas.

Conviene recordar que el objeto de estudio de esta investigación son las decisiones automatizadas que se toman a partir de la realización de inferencias algorítmicas que evalúan aspectos personales, es decir, se incluyen tanto los sistemas de apoyo a la toma de decisiones como los basados únicamente en el tratamiento automatizado.

BUSUIOC dice que los remedios y salvaguardas para la toma de decisiones algorítmica contenidas en el RGPD están restringidas a las decisiones basadas únicamente en el tratamiento automatizado[460].

A este respecto, creo oportuno realizar dos consideraciones previas que serán desarrolladas posteriormente. En primer lugar, en el Reglamento sí encontramos remedios y salvaguardas específicas para tratamientos automatizados de datos personales utilizados como apoyo a la decisión. A saber, la prohibición contenida en el 22(1) que obliga a definir la clase de intervención humana requerida para salvar dicha prohibición, así como

[460] Busuioc, «Accountable Artificial Intelligence: Holding Algorithms to Account», 7.

el 35(3)(c) cuando se refiere a decisiones basadas en el tratamiento automatizado, omitiendo el término "únicamente". Segundo, contamos con remedios y salvaguardas en el RGPD que, sin estar dirigidas a la toma de decisiones automatizada, son de extraordinaria relevancia para su aplicación, como los derechos de información y acceso sobre la elaboración de perfiles, la evaluación de impacto de protección de datos o los principios del RGPD. Todo ello será oportunamente desarrollado a lo largo de esta investigación, no obstante, conviene apreciar aquí la diferencia entre la toma de decisiones basada únicamente en el tratamiento automatizado y la elaboración de perfiles.

La elaboración de perfiles viene definida en el artículo 4 RGPD como: «*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*». Se trata de un procedimiento que, por lo general, implica una serie de deducciones estadísticas y que suele utilizarse para realizar predicciones sobre personas, utilizando datos de distintas fuentes para inferir algo sobre las mismas, sobre la base de las cualidades de otros que parecen similares estadísticamente[461]. Vemos aquí una equivalencia plena -a efectos de la presente investigación- entre la realización de inferencias algorítmicas que evalúan aspectos personales con la elaboración de perfiles definida por el RGPD[462].

En las deliberaciones legislativas, la elaboración de perfiles como tal recibió mucha más atención que la toma de decisiones automatizada[463] y, sin embargo, el RGPD confiere una protección confiere una protección menor a la elaboración de perfiles desde dos puntos

[461] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 7.

[462] Tampoco parece haber dificultades para considerar que los perfiles o inferencias algorítmicas constituyen datos personales de las personas interesadas sobre las que analizan o predicen aspectos personales, no obstante, más obstáculos encontramos a la hora de determinar el régimen jurídicos concreto y los derechos aplicables a esta clase de datos personales. Vid. Apartado 2.1. Derechos de información sobre las inferencias algorítmicas en el RGPD, en Capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica.

[463] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 84. Especialmente en lo que se refiere a los efectos discriminatorios que estos procesos podían provocar; para Veale y Edwards esta preocupación se plasma en el considerando 71, que menciona expresamente dichos efectos, y en tres áreas principales que el GT29 vincula en sus directrices a la no discriminación en la regulación de la elaboración de perfiles, Veale y Edwards, «Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling», 403.

de vista que desarrollaremos más adelante. Por un lado, si comparamos el régimen jurídico aplicable a los datos personales como datos de entrada en los modelos algorítmicos -datos personales en bruto- y como datos de salida -datos personales inferidos o perfiles de las personas interesadas-, paradójicamente estos últimos reciben menor protección en el RGPD que los primeros -o mayores limitaciones para su protección, mejor dicho-[464]. Por otro lado, la elaboración de perfiles que se utiliza como único fundamento de una decisión que afecta de forma significativa al interesado cuenta con medidas de salvaguarda específicas que, sin embargo, no se aplican a todos aquellos perfiles que no se utilizan de esta forma.

Más tarde se abordará el alcance de la referencia expresa en el artículo 22 RGPD a la elaboración de perfiles, que ha dado lugar a dos posibles interpretaciones: la primera entiende que la toma de decisiones automatizada en esta disposición incluye necesariamente elaboración de perfiles, mientras que la segunda entiende que hay lugar a decisiones automatizadas que no incluyan dicho perfilado. En cualquier caso, esta discusión no parece tener relevancia práctica ya que la mayoría de las decisiones dentro del ámbito de aplicación del artículo 22 incluirán elaboración de perfiles[465]-y así lo hace en todo caso esta investigación-.

En definitiva, encontramos en el Reglamento la figura de la elaboración de perfiles -art. 4(4)-, que puede tener como fin, o no, la toma de decisiones automatizada basada en dicho perfilado y, a su vez, esta toma de decisiones puede, o no, ser calificada como una decisión basada únicamente en el tratamiento automatizado conforme a la disposición del artículo 22 RGPD[466].

[464] Vid. más ampliamente Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI».

[465] Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 97.

[466] Conforme a la interpretación que se propone en esta investigación, veremos que la elaboración de perfiles entra en el ámbito de la toma de decisiones basada únicamente en el tratamiento automatizado cuando ésta produce un efecto jurídico o significativo en el interesado sin intervención humana significativa previa, independientemente de que dicho efecto sea abarcado o no por una voluntad "finalista" del responsable del tratamiento para la toma de decisiones.

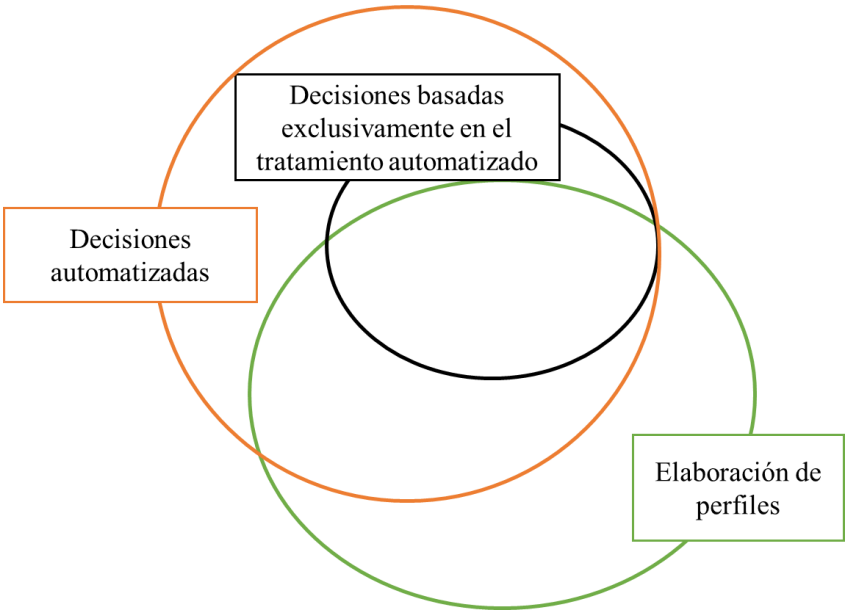

Ilustración 3. Elaborada por el autor

3.2. Dos prohibiciones y un sinfín de excepciones

La estructura del artículo 22, que comparte similitudes con su precedente como ya hemos visto, le han hecho merecedor de los sobrenombres de 'castillo de naipes' o 'rodaja de queso suizo'. El primero de ellos lo utilizó Bygrave, primero para definir a su predecesor [467], y posteriormente para refrendar que era igualmente aplicable al actual artículo 22 RGPD[468]. Para Brkan, a pesar de que, a primera vista, la prohibición de toma de decisiones automatizada parece expresar una postura reticente hacia las mismas, un análisis más profundo de la disposición saca a la luz una visión más laxa que expresa a través de la referencia a un queso suizo con gigantescos agujeros en su interior[469].

Estos sobrenombres ponen de manifiesto que, dentro de una norma sinuosa de por sí como es el RGPD, esta disposición contiene además de un número importante de excepciones a las normas generales y una ambigüedad que dificultan su análisis, interpretación y

[467] Bygrave, «Minding the machine: art 15 of the EC Data Protection Directive and automated profiling», 21.

[468] Bygrave, «Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making», 253.

[469] Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 97.

aplicación[470]. Lo que, sin duda, ha venido reforzando esa aura de derecho de segunda clase que se le ha atribuido[471].

Veamos esta estructura a la que hacen referencia estos sobrenombres. A primera vista, arroja dos prohibiciones contenidas en el apartado [1] y [4], una serie de excepciones a las mismas en los apartados [2] y [4] respectivamente, así como una serie de salvaguardas en su apartado [3], aunque también en [2.b] y [4]. Vamos a observarlo con mayor detalle.

A) Apartado primero: Todo interesado tendrá derecho a no ser objeto *(ambigüedad)* de una decisión *(amb.)* basada únicamente en el tratamiento automatizado *(amb.)*, incluida la elaboración de perfiles *(amb.)*, que produzca efectos jurídicos en él o le afecte significativamente de modo similar *(amb.)*.

Encontramos aquí un total de (5) aspectos que han sido calificados como ambiguos por la doctrina[472]. Sobre la posible resolución o no de estas ambigüedades vía interpretativa se profundizará en las respectivas secciones, si bien, corresponde enunciar brevemente en qué consisten las mismas.

En primer lugar, la redacción del derecho a no ser objeto de determinadas decisiones automatizadas da lugar a dos posibles interpretaciones, por un lado, que este derecho constituye una prohibición general para el responsable del tratamiento a adoptar decisiones basadas únicamente en el tratamiento automatizado y, por otro lado, que este derecho es en realidad una facultad a disposición de la persona interesada a ejercer o interponer a discreción de la misma[473].

---

[470] En palabras de Gil: *La gran cantidad de conceptos indeterminados permite encajar interpretaciones con diversos grados de flexibilidad o restricción que en última instancia determinará la amplitud de los derechos de los interesados y su capacidad para conocer la existencia de decisiones automatizadas, el funcionamiento de los sistemas que las toman y las garantías que pueden ejercitar para solicitar una revisión o la oposición a este tipo de decisiones*. Gil González, «Aproximación al estudio de las decisiones automatizadas en el seno del Reglamento General Europeo de Protección de Datos a la luz de las tecnologías big data y de aprendizaje computacional», 178.

[471] Bygrave, «Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making», 250.

[472] El abuso en el empleo de términos jurídicos indeterminados, así como de expresiones vagas o poco concisas es una crítica extendida al conjunto del RGPD y no solo al artículo 22. Sancho Lopez, «Legal Strategies To Ensure Fundamental Rights in Front of the Chanllenges of Big Data», 7. -si bien, esta disposición es probablemente la más críptica de toda la regulación-

[473] En esta investigación se considerará que estamos ante una prohibición general, conforme a la interpretación refrendada por el CEPD, a la aplicación observada hasta el momento en los tribunales y a la doctrina mayoritaria. Vid. Apartado 3.3. ¿Prohibición general o derecho a interponer por la persona interesada?, en este mismo capítulo.

Otro aspecto ambiguo es el relativo al alcance del término *decisión*, puesto que habrá de determinarse qué constituye cualitativamente este término entrando, por ende, en el ámbito de aplicación del artículo[474].

En tercer lugar, ha de interpretarse qué se entiende por decisiones basadas *únicamente* en el tratamiento automatizado. Aquí la discusión necesita abordar el umbral mínimo de intervención humana requerido por el propio término[475], y es que, como no podía ser de otra manera, incluir un término que se refiere al grado de automatización de un sistema requiere, inevitablemente, definir el grado de intervención humana requerida por el mismo[476].

La referencia a la inclusión de la elaboración de perfiles en las decisiones automatizadas ha dado lugar también a dos posibles interpretaciones[477]. Por un lado, se estima que la fórmula, *incluida la elaboración de perfiles*, está haciendo referencia al tratamiento automatizado y, por ende, la elaboración de perfiles es una forma de tratamiento

---

[474] Aquí se opta por considerar que el término decisión es referencial a la producción de efectos jurídicos o de afectación significativamente similar, es decir, hay decisión en tanto se producen dichos efectos y no la hay en tanto no se produzcan. Vid. Apartado 3.4. Efectos jurídicos o de afectación significativa similar: un enfoque basado en el riesgo, en este mismo capítulo.

[475] Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 244.

[476] Siendo este uno de los puntos más desarrollados en esta investigación, se considerará que ese umbral mínimo que requiere definir las decisiones basadas únicamente en el tratamiento automatizado es el de una intervención humana significativa. Esto es, aquellas decisiones que no cuenten con una intervención humana significativa habrán de ser consideradas como basadas únicamente en el tratamiento automatizado. La determinación de qué resulta significativo es, por supuesto, una tarea más compleja. Vid. Apartado 1. Derecho a la intervención humana, en capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica.

[477] Se ha señalado por Mendoza y Bygrave que esta duda interpretativa es más bien derivada de la diferencia en la redacción con el artículo 15 DPD, dado que las decisiones debían basarse *únicamente en un tratamiento automatizado de datos destinado a evaluar determinados aspectos*. Esto es, no había lugar a decisiones automatizadas que no incluyesen elaboración de perfiles. No obstante, Mendoza y Bygrave se pronuncian en sentido contrario, sobre la base de los trabajos preparatorios del RGPD entienden que no hay lugar a aplicar el artículo 22 a decisiones automatizadas que no incluyan elaboración de perfiles: *The background for Art. 22 accordingly does not support the use of the condition 'automated processing' without including profiling. Simply put, it makes most sense to treat automated processing as a condition that necessarily involves profiling. Nothing in the preamble or preparatory works runs counter to this approach. In this regard, it is worth recalling that a proposal from the European Parliament Committee on Civil Liberties, Justice and Home Affairs for regulating profiling individually without the necessity of an automated decision was expressly abandoned in favour of an approach more in line with DPD Art. 15*. Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 90-91. En sentido contrario, siguiendo las directrices del GT29, Noto La Diega entiende que la redacción del artículo 22 aporta claridad al considerar que la elaboración de perfiles no es la única forma de tratamiento automatizado que entra en su ámbito de aplicación, lo cual es una de las grandes distinciones con su precedente. Noto La Diega, «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information», 17.

automatizado que también se incluye en el ámbito de aplicación de este artículo. Por otro lado, se ha entendido que *incluida la elaboración de perfiles* estaría diferenciando entre las decisiones automatizadas y la elaboración de perfiles, siendo un término independiente del otro[478].

Por último, la producción de efectos jurídicos o de afectación significativamente similar es fuente de ambigüedad. Aquí se introduce un enfoque basado en el riesgo, en virtud del cual, ha de determinarse el umbral mínimo de riesgo que debe producirse para determinar la aplicación o no de la prohibición de esta disposición. No exenta de matices, la determinación de qué se consideran efectos jurídicos, puede ofrecer una necesaria seguridad jurídica. Sin embargo, los efectos que afectan a la persona interesada *significativamente de modo similar*, cuya redacción de por sí es confusa, puede dar lugar a múltiples interpretaciones[479].

B) Apartado segundo: 2. El apartado 1 no se aplicará si la decisión: a) es necesaria para la celebración o la ejecución de un contrato entre el interesado y un responsable del tratamiento *(excepción)*; b) está autorizada por el Derecho de la Unión o de los Estados miembros que se aplique al responsable del tratamiento *(exc.)* y que establezca asimismo medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado, o; c) se basa en el consentimiento explícito del interesado *(exc.)*.

En el segundo apartado encontramos tres excepciones a la que, como hemos adelantado, consideramos una prohibición general establecida en el primer apartado. Es decir, los responsables del tratamiento podrán adoptar decisiones basadas únicamente en el tratamiento automatizado siempre que cumplan con alguna de las excepciones o bases legitimadoras contenidas en este segundo apartado. Excepciones basadas en la necesidad

[478] No teniendo gran repercusión esta diferencia interpretativa, en esta investigación se opta por la primera. Es decir, la elaboración de perfiles es una forma de tratamiento automatizado que se incluye específicamente en el ámbito de aplicación de este artículo. Conforme a la redacción del Considerando 71 y a la definición misma de la elaboración de perfiles en el artículo 4(4), parece la interpretación más apropiada.

[479] Como punto de partida, siguiendo las directrices del GT29, en esta investigación se entiende que la inclusión del término similar hace que la determinación de la significancia se vincule al umbral producido por un efecto jurídico, en línea con Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 21. Ahora, no se cierra así debate alguno dados los diferentes efectos que pueden producir distintos actos jurídicos. Vid. Apartado 3.4. Efectos jurídicos o de afectación significativa similar: un enfoque basado en el riesgo, en este mismo capítulo.

contractual[480], en el consentimiento de la persona interesada[481] y en el Derecho de la UE y los EEMM[482].

Ahora bien, a la aplicación de estas excepciones les sigue, en todo caso, la condición de aplicar una serie de medidas para salvaguardar los derechos y libertades de las personas interesadas. En el caso de las excepciones relativas a la necesidad contractual y al consentimiento del interesado, estas salvaguardas se encuentran en el apartado tercero que abordaremos a continuación. Mientras que, cuando la excepción venga dada por el Derecho de la Unión o de los EEMM, será dicha normativa la que establezca las salvaguardas necesarias[483].

[480] En cuanto a la amplia gama de excepciones contenidas en el artículo 22, Hilden afirma que el grado de aceptabilidad del procesamiento de datos en el RGPD se determina por lo que se percibe como innovador por el legislador y, por lo tanto, el papel de los datos y las decisiones automatizadas no se cuestiona profundamente: *The exceptions to Article 22 on automatic decision-making are especially informative in this regard, as it is blatantly obvious that industries most reliant on automatic processing, such as the financial and insurance industries, are exempt based on contractual necessity*. Hilden, «The Politics of Datafication: The influence of lobbyists on the EU's data protection reform and its consequences for the legitimacy of the General Data Protection Regulation», 198. Según Roig, la amplitud de las excepciones muta la regla general y le dan a la prohibición del apartado primero, en cambio, un carácter residual, en Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 39.

[481] Son también destacables las críticas que ha recibido el consentimiento como base legitimadora para la toma de decisiones automatizada, excediendo las mismas el objeto de esta investigación, se recomiendan los textos de Gil y De Hert y Noto La Diega. En ambos se destaca el papel de la información para el ejercicio de dicho consentimiento por el interesado. Los primeros critican especialmente la ambigüedad contenida en los derechos de información del RGPD en relación con la toma de decisiones automatizada, Gil González y Hert, «Understanding the legal provisions that allow processing and profiling of personal data—an analysis of GDPR provisions and principles», 615. Mientras que el segundo destaca el desequilibrio de poder entre los responsables del tratamiento (utiliza un banco como ejemplo) y las personas interesadas (un cliente que quiere acceder a un crédito) y su impacto en el libre ejercicio del consentimiento, en Noto La Diega, «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information», 20. También sobre el consentimiento expreso, vid. Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 41 y ss. Y más ampliamente, sobre cómo los sistemas de IA han afectado al rol del consentimiento en la protección de datos, vid. Jones y Edenberg, «Troubleshooting AI and Consent».

[482] Al contrario que en el apartado 4, donde esta misma cláusula sí se limita a fines de interés público esencial, esta excepción no contiene aparentes límites. Esto, en principio, podría llevarnos a pensar en escenarios en los que los Estados utilizasen esta excepción para la legitimación, por ejemplo, de sistemas de vigilancia masiva. Noto La Diega entiende que el hecho de que la excepción deba estar basada en la ley y la referencia al Derecho de la UE, implicaría que debe interpretarse como una facultad para legislar más allá de lo ya previsto por los tratados, en Noto La Diega, «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information», 21.

[483] Esta cláusula ha dado lugar a diferentes interpretaciones por parte de los EEMM. No solo en lo relativo al desarrollo y aplicación de distintas medidas para salvaguardar los derechos y libertades de los interesados, sino también a la hora de interpretar en qué medida esta cláusula permite ampliar la propia prohibición de la toma de decisiones automatizada, tal y como se apunta en la anterior nota al pie.

C) Apartado tercero: En los casos a que se refiere el apartado 2, letras a) y c), el responsable del tratamiento adoptará las medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado, como mínimo el derecho a obtener intervención humana por parte del responsable, a expresar su punto de vista y a impugnar la decisión.

En este apartado se enumeran las medidas mínimas para salvaguardar los derechos y libertades de las personas interesadas cuando se adoptan decisiones automatizadas sobre la base legitimadora de necesidad contractual o del consentimiento. Son tres las medidas que el interesado debe estar en disposición de ejercer tras la toma de la decisión automatizada: derecho a obtener intervención humana[484], derecho del interesado a expresar su punto de vista y a impugnar la decisión[485]. No puede obviarse la medida ausente en este apartado que ha suscitado un intenso debate doctrinal: el derecho a una explicación. Este derecho está expresamente mencionado en el considerando 71, pero no tuvo su reflejo en la redacción final del apartado 3[486].

D) Apartado cuarto: Las decisiones a que se refiere el apartado 2 no se basarán en las categorías especiales de datos personales contempladas en el artículo 9, apartado 1, salvo que se aplique el artículo 9, apartado 2, letra a) o g) *(exc. x2)*, y se hayan tomado medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado.

[484] Siendo muy relevante distinguir entre este mecanismo de intervención humana – derecho a obtenerla tras la toma de una decisión basada únicamente en el tratamiento automatizado – y el contenido en el apartado primero dada la prohibición de toma decisiones basadas únicamente en el tratamiento automatizado – es decir, la obligación del responsable de tratamiento a introducir intervención humana de forma previa a la toma de decisiones para esquivar la prohibición –. Vid. Apartado 1. Derecho a la intervención humana, en Capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica.

[485] Tal y como se desarrollará más adelante, siguiendo la argumentación de Bayamlıoğlu, las medidas de intervención humana y a expresar su punto de vista son auxiliares a la columna vertebral de las mismas, esto es, el derecho a contestar la decisión. Bayamlıoğlu, «The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"», 5.

[486] Esta decisión del legislador permite distintas lecturas que han dado lugar a ese intenso debate que se abordará más adelante. Vid. Apartado 2.2. Derecho a la información para las decisiones basadas únicamente en el tratamiento automatizado, ¿derecho a una explicación?, en Capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica.

Por último, el artículo 22 incluye una segunda prohibición general[487] para aquéllas decisiones basadas únicamente en el tratamiento automatizado que, encontrándose en principio legitimadas por una de las excepciones del apartado segundo, se basen en las categorías especiales recogidas en el artículo 9 RGPD, esto es: «*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 las orientación sexuales de una persona física*».

Ahora bien, una vez más el Reglamento establece excepciones a esta prohibición general, basadas o bien consentimiento explícito de la persona interesada[488], o bien cuando sobre la base del Derecho de la Unión o de los EEMM la decisión sea necesaria por razones de un interés público esencial[489]. A dichas excepciones se añaden igualmente la condición de establecer una serie de salvaguardas para los derechos y libertades de las personas interesadas[490]. Un último apartado en el que se refleja la preocupación del legislador por la toma de decisiones basadas en dichas categorías especiales a partir de unas restricciones específicas pero que, bien es cierto, no se traduce en medidas de salvaguarda más gravosas[491].

3.3. ¿Prohibición general o derecho a interponer por la persona interesada?

[487] Sobre esta prohibición sí que parece no haber dudas sobre su naturaleza, esto es, no necesita ser invocada por el interesado, así lo admite Tosoni, «The right to object to automated individual decisions: resolving the ambiguity of Article 22(1) of the General Data Protection Regulation», 12.

[488] Excepto cuando el Derecho de la Unión o de los EEMM establezca que la prohibición no puede ser levantada por el interesado -art. 9(2)(a) RGPD-.

[489] Debiendo ésta ser proporcional al objetivo perseguido, respetar en lo esencial el derecho a la protección de datos y establecer medidas adecuadas y específicas para proteger los intereses y derechos fundamentales del interesado -art. 9(2)(g) RGPD-.

[490] Sin embargo, un aspecto ambiguo a este respecto es el de qué clase de salvaguardas deben establecerse para cada excepción. En relación con la excepción sobre el Derecho de la Unión o de los EEMM, parece claro que será la misma normativa que legitima el tratamiento de forma excepcional la que establezca dichas medidas, en analogía con lo establecido para la excepción del 22(2)(b). Sin embargo, en relación con el consentimiento explícito no queda claro si irremediablemente hemos de acudir a las establecidas de forma genérica, y también de forma análoga para el 22(2)(c), en el apartado tercero del artículo 22.

[491] Korff propuso que los perfiles que puedan discriminar sobre las categorías de datos de carácter especial deben declararse expresamente contrarios al RGPD: *the Regulation should expressly stipulate that profiles that (whether intentionally or otherwise) have the effect of discriminating against individuals on the basis of race or ethnic origin, political opinions, religion or beliefs, trade union membership, or sexual orientation, shall be contrary to the Regulation*. Korff, «Comments on Selected Topics in the Draft EU Data Protection Regulation», 35.

A pesar de que, como se verá a continuación, el CEPD, la jurisprudencia y la mayoría de la doctrina coinciden en que el derecho enunciado en el párrafo primero del artículo 22 constituye una prohibición general a adoptar decisiones basadas únicamente en el tratamiento automatizado[492], hay parte de la doctrina que defiende que este derecho es, en realidad, una facultad a disposición de la persona interesada a ejercer o interponer a discreción de la misma[493].

Es decir, bajo la primera opción (prohibición general), el responsable del tratamiento no estaría autorizado a adoptar decisiones basadas únicamente en el tratamiento automatizado salvo que cumpla con alguna de las excepciones contenidas en el párrafo segundo, o un agente humano intervenga en la toma de decisiones evitando que ésta se base únicamente en dicho tratamiento automatizado.

Al contrario, bajo la segunda opción (derecho de oposición), el responsable podría adoptar esta clase de decisiones siempre que la persona interesada no se oponga de forma activa la decisión que le afecte generándole un efecto jurídico o significativamente similar.

Aparentemente, el debate no está cerrado[494]. Resolver esta ambigüedad es un punto crítico para esta disposición, dadas las diferentes consecuencias que tienen una y otra posición

[492] Las directrices del GT29 refrendadas por el CEPD son claras al respecto: "El artículo 22, apartado 1, establece una prohibición general de las decisiones basadas únicamente en el tratamiento automatizado. Esta prohibición se aplica tanto si el interesado adopta una acción relativa al tratamiento de sus datos personales como si no lo hace.", vid. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 21. (GT29, 21). Entre quienes se han manifestado a favor de esta interpretación mayoritaria, Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -»; Jones, «The right to a human in the loop: Political constructions of computer automation and personhood»; Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond»; Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*; Sartor y Lagioia, «The impact of the General Data Protection Regulation (GDPR) on artificial intelligence (PE 641.530)».

[493] Lamentablemente, es también la posición que parece adoptar la LOPDGDD, cuando incluye un derecho de oposición a las decisiones individuales automatizadas entre los deberes de transparencia e información para ejercer los derechos 15 a 22 RGPD. Artículo 11 LOPDGDD: *(…) Si los datos obtenidos del afectado fueran a ser tratados para la elaboración de perfiles, la información básica comprenderá asimismo esta circunstancia. En este caso, el afectado deberá ser informado de su derecho a oponerse a la adopción de decisiones individuales automatizadas que produzcan efectos jurídicos sobre él o le afecten significativamente de modo similar, cuando concurra este derecho de acuerdo con lo previsto en el artículo 22 del Reglamento (UE) 2016/679*.

[494] Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 92.

para los responsables y las personas interesadas[495]. En este apartado se aborda este debate, tratando de contrarrestar los argumentos esgrimidos por la posición minoritaria para mostrar la conveniencia de interpretar el derecho a no ser objeto de decisiones automatizadas como una prohibición general.

En relación con dicha posición minoritaria, ha de reconocerse la extraordinaria labor argumentativa desarrollada por TOSONI[496]. La lectura de su artículo merece ser tomada como referencia de dicha posición.

En primer lugar, TOSONI hace referencia a una interpretación gramatical partiendo de varias premisas que considero plenamente válidas; que el TJUE tiende a rechazar interpretaciones que se alejen del contenido literal de la disposición; que las distintas traducciones oficiales de la disposición arrojan ligeras variaciones terminológicas; y que los términos que merecen un especial escrutinio son "derecho", "ser objeto" y "decisión"[497]. Hasta este punto, las premisas de las que parte el autor no permiten decantar la balanza hacia ninguno de los dos lados, como él mismo reconoce. Sin embargo, hay otros argumentos que deben ser objeto de crítica en su lectura literal.

Primero, se refiere a la distinción realizada en el título de la sección 4ª del capítulo 3º del RGPD: "Derecho de oposición y decisiones individuales automatizadas"[498]. Este título según el GT29 implica que el artículo 22 no es derecho de oposición como el artículo 21[499], por el contrario, TOSONI indica que bajo este título se engloban un derecho de

[495] Wachter, Mittelstadt, y Floridi, «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation», 95.

[496] En España ha mantenido esta posición Guzmán Fluja en contraposición con la prohibición contenida en el análogo artículo 11 de la Directiva 680/2016, sin embargo, simplemente sostiene esta posición a partir de la distinta redacción de una y otra disposición. Vid. Guzman Fluja, «Proceso penal y justicia automatizada».

[497] Tosoni, «The right to object to automated individual decisions: resolving the ambiguity of Article 22(1) of the General Data Protection Regulation», 8-9.

[498] También se ha esgrimido que, dado que el artículo 22 fue incluido en el capítulo III sobre derechos del interesado, el legislador no quiso concebirlo como una prohibición. Sin embargo, ya hemos dicho que el término "derecho" no debe ser necesariamente entendido como una facultad a disposición del interesado, y el propio Roig reconoce después que otros artículos contenidos en dicho capítulo, cita el 13 y el 14, se configuran como deberes de los responsables del tratamiento. Vid. Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 92.

[499] Esto se pone todavía más de relieve por la ausencia de una obligación de información explícita correspondiente en el artículo 22, que sí aparece en el artículo 21, apartado 4, en Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y

oposición al tratamiento y un derecho de oposición a la toma de decisiones[500]. A mi entender, esta última lectura a partir de dicho título es, sin duda, forzada en términos gramaticales, tanto en lengua inglesa como castellana[501].

Segundo, hace referencia al argumento del GT29 en relación con el considerando 71 y las excepciones contenidas en el artículo 22.2. Las directrices entienden que los términos "se deben permitir" en dicho considerando como que hay un levantamiento de la prohibición en el apartado segundo[502]. Sin embargo, TOSONI hace referencia a las versiones italiana y neerlandesa para defender que dichos términos se entienden de forma inequívoca por razón de oportunidad[503]. Ahora bien, las versiones en inglés o castellano arrojan términos igualmente inequívocos en el sentido contrario recogido en las directrices, con lo cual, sería razonable defender que ambas lecturas caben en este caso bajo una interpretación gramatical.

En segundo lugar, Tosoni alega que hay una "lógica interna" del artículo 22 que invita a entender el apartado primero como una facultad a disposición del interesado. Esta posición la fundamenta en el consentimiento recogido en el apartado segundo y la prohibición de tratamiento de categorías esenciales del apartado cuarto[504]. En lo referido a la prohibición del apartado cuarto -que sí reconoce como prohibición-, Tosoni entiende que la referencia de dicho apartado a las decisiones "a que se refiere el apartado 2", no implica que la prohibición no se aplique también a las decisiones que, según su

elaboración de perfiles a los efectos del Reglamento 2016/679», 39. Esta diferencia también se hace explícita al comparar los Considerandos 70 y 71.

[500] Tosoni, «The right to object to automated individual decisions: resolving the ambiguity of Article 22(1) of the General Data Protection Regulation», 10.

[501] No obstante, el legislador español parece haber interpretado el artículo 22 como un derecho a oposición en el artículo 11 de la LOPDGDD: 2. (…) *Si los datos obtenidos del afectado fueran a ser tratados para la elaboración de perfiles, la información básica comprenderá asimismo esta circunstancia. En este caso, el afectado deberá ser informado de su derecho a oponerse a la adopción de decisiones individuales automatizadas que produzcan efectos jurídicos sobre él o le afecten significativamente de modo similar, cuando concurra este derecho de acuerdo con lo previsto en el artículo 22 del Reglamento (UE) 2016/679*. Siendo el RGPD directamente aplicable en nuestro ordenamiento no parece que esta lectura tenga excesiva relevancia, más cuando termina referenciando lo previsto en el artículo 22 RGPD.

[502] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 40.

[503] Tosoni, «The right to object to automated individual decisions: resolving the ambiguity of Article 22(1) of the General Data Protection Regulation», 10.

[504] Tosoni, 11-12. Acerca de la lógica sobre el consentimiento que argumenta, entiendo que la misma se fundamenta sobre un entendimiento erróneo de la relación entre el término "decisión" y la producción de efectos jurídicos y significativamente similares, así como de obviar la intervención humana contenida en el apartado primero, con lo cual esta cuestión será resuelta en las secciones pertinentes.

interpretación, son legítimas bajo el paraguas del apartado primero –si bien oponibles por la persona interesada–[505]. Ahora bien, para ello realiza un salto que tiene difícil justificación a nivel gramatical y que, sin embargo, encaja perfectamente con la interpretación del apartado primero como una prohibición; esto es, el apartado primero prohíbe con carácter general determinadas decisiones automatizadas y el apartado segundo establece una serie de excepciones que, no obstante, no pueden estar basadas en el tratamiento de categorías especiales por la prohibición del apartado cuarto –salvo que, una vez más, dicho tratamiento se ampare en alguna de las nuevas excepciones contenidas en el mismo–.

En relación con la lógica interna de la disposición, cabe traer aquí a colación el argumento esgrimido por Mendoza y Bygrave, con el que coincido plenamente, cuando señalan que bajo la interpretación minoritaria el apartado primero funcionaría como un derecho a exigir intervención humana en la decisión pertinente -basada únicamente en el tratamiento automatizado-, lo cual convertiría en superflua la salvaguarda del apartado 3 también referida a la intervención humana[506].

Asimismo, Tosoni se apoya en una interpretación teleológica para exponer que el artículo 22(1) sirve mejor a los objetivos que persigue como un derecho a objetar la decisión automatizada. Entiende que el objetivo fundamental del mismo, en cumplimiento del considerando sexto, es el de garantizar un elevado nivel de protección de los datos personales[507]. Así, esgrime que el Convenio 108+ no necesita establecer una prohibición

[505] Tosoni, 12.

[506] Para ellos: *Así pues, tanto desde un punto de vista lógico como desde una perspectiva más teleológica basada en la preocupación por la privacidad y la protección de datos como derechos fundamentales, lo más lógico es concluir que el aparente derecho previsto en el art. 22(1) no tiene que ser ejercido por el interesado*. [Thus, from both a logical point of view and a more teleological perspective rooted in concern for privacy and data protection as fundamental rights, it makes most sense to conclude that the apparent right provided by Art. 22(1) does not have to be exercised by the data subject]. Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 87. Esta contradicción se sumaría a la señalada por el GT29: *Si el artículo 22 se interpretara como un derecho de oposición, la excepción de su apartado 2, letra c), no tendría mucho sentido. La excepción establece que las decisiones automatizadas se siguen pudiendo adoptar si el interesado ha dado su consentimiento explícito (véase más abajo). Esto sería contradictorio, ya que el interesado no puede consentir y oponerse a un mismo tratamiento*. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 39.

[507] Para ello hace referencia al fundamento de esta disposición, tal y como lo desarrollan Mendoza y Bygrave, cuestión que desarrollaré en mayor profundidad para hablar del fundamento de la intervención humana. Vid. Apartado 1.2.2. Fundamento para la intervención humana en el artículo 22 RGPD, en Capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica.

general para garantizar un alto nivel de protección *en línea con el RGPD* y que, estando los objetivos del artículo 22 en línea con su predecesor el artículo 15 DPD, debería ser suficiente con continuar el enfoque del mismo para cumplir dichos objetivos[508].

Dos objeciones a estos argumentos. Por un lado, el hecho de que el renovado Convenio 108+ se declare en línea con el RGPD, no implica que uno y otro texto deban coincidir en todos los aspectos, ni que ello resulte apropiado[509]. Por otro lado, el fracaso ya constatado del artículo 15 DPD debería ser suficiente para enfatizar los cambios que ofrece esta nueva versión, siendo conscientes de que el legislador tenía por objetivo mejorar la regulación existente[510]. Tal y como sostiene Binns, en un argumento sobre el que volveremos más adelante, hacer depender la aplicación de este mecanismo en la objeción individual puede socavar la aplicación equitativa de la justicia individual[511].

En definitiva, en este caso, no parece razonable argüir que el artículo 22 ofrece un nivel más *alto* de protección para las personas interesadas concebido como un derecho a objetar[512]. Cosa distinta es que se considere que, de este modo, el artículo 22 ofrece una protección más adecuada en el ecosistema regulatorio del RGPD. En este sentido, Bygrave entiende que un derecho de oposición, primero, reconocería el hecho de que estos procesos de decisión ya se utilizan ampliamente tanto en el sector privado como en

[508] Tosoni, «The right to object to automated individual decisions: resolving the ambiguity of Article 22(1) of the General Data Protection Regulation», 17.

[509] Ello podría resultar contraproducente para que terceros estados se adhieran a la regulación del Convenio, privando a éste de adquirir un modelo más flexible que el RGPD para dichos fines, según de Hert y Papakonstantinou, «Framing Big Data in the Council of Europe and the EU data protection law systems: Adding 'should' to 'must' via soft law to address more than only individual harms», 640. Adquiere, por tanto, pleno sentido que el RGPD ofrezca un modelo más estricto (prohibición general) que el del Convenio 108+ (derecho a objetar).

[510] Con mayor énfasis, pues, en aquellos aspectos en los que la Directiva 95/46/CE se había mostrado inútil.

[511] En sus propias palabras: *(…), it puts the onus on decision subjects to mount a successful challenge, which may require resources and privileges that are not distributed equally, compounding disadvantage by disproportionately preserving individual justice for those who are already advantaged*. Binns, «Human Judgment in algorithmic loops: Individual justice and automated decision-making», 11.

[512] Para el GT29: *Esta interpretación refuerza la idea de que sea el interesado quien tenga el control sobre sus datos personales, lo cual se corresponde con los principios fundamentales del RGPD. Interpretar el artículo 22 como una prohibición en vez de como un derecho que debe invocarse significa que las personas están protegidas automáticamente frente a las posibles consecuencias que pueda tener este tipo de tratamiento*. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 22.

el público, en ámbitos en los que la digitalización está avanzada[513]. Segundo, reconocería que estos procesos de decisión pueden tener beneficios socialmente justificables[514].

Dada la línea argumental que se reitera en esta investigación, no pueden tomarse por válidos dichos argumentos. En primer lugar, la utilización masiva de esta clase de toma de decisiones no legitima las mismas, sino que es una muestra del desequilibrio de poder y el fracaso de las herramientas para su control que, hasta el momento, ha impuesto el Derecho. En segundo lugar, la concepción del artículo 22 como una prohibición general no implica que dichos procesos puedan tener beneficios socialmente. Implica, sin embargo, que si un responsable del tratamiento no cumple con alguna de las -amplias- excepciones de la prohibición y, por consiguiente, con las medidas de salvaguarda debidas, está obligado, al menos, a introducir una intervención humana significativa en este proceso para evitar la prohibición.

La introducción de salvaguardas en la toma de decisiones automatizada no implica un desvalor sobre las mismas, sino la conciencia sobre los posibles riesgos asociados a sus potenciales beneficios. Además, y por último, un derecho de oposición como el configurado en el artículo 21 a disposición del interesado, significaría que éste puede denegar los efectos de una decisión automatizada independientemente de si esta decisión es lícita y exacta. Esta visión, que hace depender la vigencia de las decisiones automatizadas de la voluntad de un sujeto, cuestiona mucho más los beneficios de estas decisiones que una prohibición general que admita excepciones sobre la base de medidas de salvaguarda de los derechos y libertades.

Habiendo discutido los argumentos más relevantes esgrimidos por la posición minoritaria y aunque algunos de ellos se discuten también en próximas secciones con mayor precisión, es conveniente advertir qué posición se ha adoptado por la escasa, hasta el momento, jurisprudencia de la que disponemos. Tanto en el caso SyRI resuelto por el

[513] Incluso se ha considerado que la interpretación del 22(1) supondría limitar más allá de lo razonable el principio de libertad de empresa, en Armada Villaverde y López Bustabad, «El reglamento general de protección de datos ante el fenómeno del "«big data"». Este argumento parece obviar el mero hecho de que la prohibición simplemente obliga, para aquellos casos en los que no se cumpla con alguna de las bases legitimadoras del apartado segundo, a que dicho tratamiento no se base únicamente en el tratamiento automatizado. Hablar de una limitación más allá de lo razonable a la libertad de empresa por exigir un determinado grado de intervención humana en el proceso decisorio parece desorbitado.

[514] Bygrave, «Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making», 253.

Tribunal de Distrito de La Haya, como en los casos Uber y Ola resueltos por el Tribunal de Distrito de Ámsterdam, se han seguido en este aspecto las directrices refrendadas por el CEPD.

Respecto del primero, a pesar de que la aplicación del artículo 22 no formaba parte del grueso de la sentencia, el Tribunal de Distrito de La Haya puntualizó en 2020 algunos aspectos sobre el mismo en relación con el caso. Entre dichos aspectos, el tribunal destacó que: «*En virtud del artículo 22 RGPD, se prohíbe en general la toma de decisiones totalmente automatizada, incluida la elaboración de perfiles, que produzcan efectos jurídicos o afecten significativamente al interesado*»[515]. Siguiendo esta misma línea en 2021, y esta vez sí en tres casos en los que se aplicó directamente el artículo 22, el Tribunal de Distrito de Ámsterdam dejaba poco margen a la interpretación, declarando la existencia de una prohibición general en el apartado primero de esta disposición[516].

Por tanto, teniendo en cuenta el *soft law* refrendado por el CEPD, la aplicación realizada por distintos tribunales y los argumentos doctrinales esgrimidos anteriormente, entiendo que queda poco espacio ya para la ambigüedad que inicialmente podía suscitar este debate dada la confusa redacción que arroja este Artículo 22. En definitiva, el apartado primero debe ser calificado y entendido como una prohibición general de adoptar decisiones basadas únicamente en el tratamiento automatizado, en virtud de la cual el responsable del tratamiento está sujeto a ella sea o no invocada por la persona interesada.

3.4. Efectos jurídicos o de afectación significativa similar: un enfoque basado en el riesgo

Como ya hemos adelantado, la producción de efectos jurídicos o de afectación significativa similar por el tratamiento automatizado de datos, incluida la elaboración de perfiles, determina la aplicabilidad de este artículo.

---

[515] Sentencia del Tribunal de Distrito de La Haya [Rechtbank Den Haag] de 5 de febrero de 2020 (ECLI: NL: RBDHA:2020: 865), *SyRI case* (par. 6.35). [*Op grond van artikel 22 AVG geldt een algemeen verbod op volledig geautomatiseerde individuele besluitvorming, met inbegrip van profilering, waaraan voor de betrokkene juridische of anderszins aanmerkelijke gevolgen zijn verbonden*]. Traducción realizada a partir del traductor automático *DeepL*.

[516] Sentencias del Tribunal de Distrito de Ámsterdam [Rechtbank Amsterdam] de 11 de marzo de 2021; *Uber deactivation case* -par. 4.8- (ECLI:NL:RBAMS:2021:1018); *Ola transparency case* -par. 4.51- (ECLI:NL:RBAMS:2021:1019); y *Uber transparency request case* -par. 4.66- (ECLI:NL:RBAMS:2021:1020). También se ve reflejada la existencia de esa prohibición general en los 3 resúmenes aportados por el Tribunal de Distrito para cada una de las sentencias.

Para BRKAN, el artículo 22 no se aplica a decisiones de bajo impacto[517]. En este mismo sentido, sostengo que el artículo 22 está cimentado sobre un enfoque basado en el riesgo que tiene una doble repercusión en su aplicación. Por un lado, la toma de una decisión automatizada –basada o no únicamente en el tratamiento– depende de si el tratamiento supera ese umbral determinado por la producción de dichos efectos[518]. Por otro lado, a la hora de discernir si determinado tratamiento entra en el ámbito de aplicación de esta disposición, al responsable del tratamiento le corresponde determinar, en primer lugar, si dicho tratamiento produce o no estos efectos. Antes de observar con detenimiento qué implican estos dos aspectos, veamos por qué el artículo 22 se cimenta sobre un enfoque basado en el riesgo.

Según GELLERT, el enforque basado en el riesgo[519], en vez de aplicar los distintos principios de protección de datos de forma equitativa, proporciona una protección que es, por definición, desigual en la medida en que depende directamente del nivel de riesgo en juego para cada operación de tratamiento específica[520]. Es decir, primero se debe evaluar dicho riesgo y, en consecuencia, determinar el régimen jurídico aplicable a la operación en cuestión.

Este enfoque está, a su vez, íntimamente unido al principio de responsabilidad en el RGPD, dado que el objetivo del mismo es lograr el cumplimiento efectivo de las normas de protección de datos, a la luz de los nuevos y mayores riesgos que plantean las nuevas tecnologías[521]. El alcance de las obligaciones jurídicas de los responsables del tratamiento depende del riesgo que planteen sus operaciones de tratamiento[522]. En definitiva, el

[517] Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 93. Solo a efectos de *graves consecuencias* según Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 23.

[518] Efectos jurídicos o significativos, en adelante.

[519] Este enfoque está particularmente extendido en el contexto normativo europeo y, lo que resulta más relevante respecto del objeto de esta investigación, ocupa un lugar destacado en las propuestas de regulación de los sistemas de IA, vid. Apartado 3.1. Propuestas para la regulación europea de la inteligencia artificial, en Introducción a la gobernanza y supervisión humana de la toma de decisiones automatizada basada en la elaboración de perfiles.

[520] Gellert, «Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms?», 3.

[521] Demetzou, «Data Protection Impact Assessment: A tool for accountability and the unclarified concept of 'high risk' in the General Data Protection Regulation», 4.

[522] Vid. más ampliamente sobre la gestión del riesgo en el RGPD el Apartado 2. Enfoque basado en la evidencia en el RGPD. Principio de responsabilidad -*accountability*-: desplazando la carga desde el individuo hacia el responsable del tratamiento, en el Capítulo 4. La intervención humana y el principio de

RGPD establece una graduación sobre las obligaciones jurídicas basada en el riesgo directamente relacionada con la responsabilidad de cumplimiento de las mismas y su demostración[523].

El hecho de que no aparezca explícitamente el término "riesgo", tal y como lo hace en el artículo 35 para la evaluación de impacto de protección de datos[524], no quiere decir que este enfoque basado en el riesgo no esté presente en este artículo 22. Es más, podríamos decir que en cierto sentido hay una vinculación entre un enfoque y el otro, dado que ha de considerarse como de "alto riesgo", en todo caso: «*la evaluación sistemática y exhaustiva de aspectos personales de personas físicas que se base en un tratamiento automatizado, como la elaboración de perfiles, y sobre cuya base se tomen decisiones que produzcan efectos jurídicos para "las persona" físicas o que les afecten significativamente de modo similar*»-art. 35(3)(a) RGPD-. Es decir, la toma de decisiones que produce esta clase de efectos representa un riesgo[525] que se ve agravado –hasta el punto de calificarse de alto riesgo– cuando se realiza a partir de dichas evaluaciones sistemáticas y exhaustivas de personas físicas.

Un segundo aspecto que vincula la determinación de los efectos jurídicos o significativos bajo el artículo 22 y el alto riesgo para los derechos y libertades fundamentales en el artículo 35 es la vulnerabilidad[526]. Esta conexión, identificada por Malgieri y Niklas, se

responsabilidad en el tratamiento de datos personales: un enfoque basado en la evidencia a través de la evaluación de impacto. Una propuesta desde la medicina preventiva.

[523] Y lo hace porque es la producción de estos riesgos lo que determina una mayor o menor intrusión en el derecho a la vida privada de las personas interesadas. En este sentido, el Tribunal de Distrito de La Haya en el caso SyRI argumentaba que, en dicho caso, no era relevante el hecho de si el tratamiento en cuestión podía encuadrarse en la toma de decisiones automatizada del artículo 22, pero sí lo era el hecho de que el tratamiento produjese los efectos descritos, dado que determina, en parte, la extensión en la que la legislación sobre SyRI interfiere en el derecho a la vida privada y familiar reconocida en el artículo 8 CEDH (par. 6.60). Sentencia del Tribunal de Distrito de La Haya [Rechtbank Den Haag] de 5 de febrero de 2020 (ECLI: NL: RBDHA:2020: 865).

[524] La inclusión explícita del término "alto riesgo" no quiere decir que, por ello, el enfoque basado en el riesgo contenido en el artículo 35 sea más claro que el contenido en el artículo 22 y la "producción de efectos jurídicos o de afectación significativa". La ambigüedad de dicho "alto riesgo" ha sido también motivo de crítica, vid. Yeung y Bygrave, «Demystifying the modernized European data protection regime: Cross-disciplinary insights from legal and regulatory governance scholarship».

[525] La toma de decisiones automatizada con efecto jurídico significativo o similar – basada o no únicamente en el tratamiento automatizada – es, por sí, uno de los nueve criterios a considerar por entrañar un probable alto riesgo, vid. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «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», 10.

[526] Malgieri y Niklas, «Vulnerable data subjects», 7.

observa en las distintas directrices del GT29, así, la vulnerabilidad de las personas interesadas puede determinar que se produzcan efectos significativos en determinado tratamiento[527], y también la vulnerabilidad es uno de los nueve criterios a considerar por entrañar un probable alto riesgo[528].

En definitiva, bajo el ecosistema regulatorio del RGPD, la toma de decisiones automatizada que produzca determinados efectos es, en sí, un riesgo; asimismo, a la hora de determinar la producción de esos "efectos", hay criterios que coinciden con la determinación de un "alto riesgo". Por ende, sin dejar de ser un derecho de las personas interesadas, el artículo 22 se aplica sobre un enfoque basado en el riesgo dada la necesidad de determinar la producción de efectos jurídicos o significativos para su aplicación, constituyendo el artículo 22 una obligación jurídica para el responsable del tratamiento dependiendo del riesgo que planteen sus operaciones.

Tal y como se ha adelantado, ello se traduce, por un lado, en que la toma de una decisión automatizada –basada o no únicamente en el tratamiento– depende de si el tratamiento supera ese umbral determinado por la producción de dichos efectos. Es decir, la producción de un efecto es la condición que determina la toma de una decisión. Así, no es necesario que el responsable adopte una decisión en sentido finalista, sino que basta que el tratamiento automatizado, incluida la elaboración de perfiles, supere ese umbral de riesgo determinado por la producción de un efecto jurídico o de afectación significativa similar. En este sentido se ha discutido si una etapa intermedia o puntual que se realiza durante el tratamiento automatizado puede o no constituir una decisión automatizada. Noto La Diega sugiere que en muy raras ocasiones esta clase de medidas podrán acogerse a la aplicación del artículo 22 RGPD[529], sin embargo, para Kamarinou et al. el

[527] Para el GT29: *Un tratamiento que pueda tener poco impacto sobre las personas en general podría tener de hecho un efecto significativo en determinados grupos de la sociedad, como grupos minoritarios o adultos vulnerables*. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 22.

[528] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «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», 11. También en el Considerando 75 RGPD: *Los 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 se traten datos personales de personas vulnerables, en particular niños*.

[529] Noto La Diega, «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information», 18.

Reglamento permite realizar una interpretación amplia, incluyendo etapas intermedias o puntuales, a partir del contenido del Considerando 71: «*El interesado debe tener derecho a no ser objeto de una decisión, que puede incluir una medida*»[530].

A pesar de que esta cuestión no se aborda directamente por las directrices refrendadas por el CEPD, ello no ha impedido que la jurisprudencia se haya manifestado en favor de esa segunda posición.

En relación al caso SyRI, el Tribunal de Distrito de La Haya argumenta que un informe de riesgo genera por sí un efecto que afecta de manera significativa al interesado, aunque dicho informe no necesariamente conlleve una investigación o sanción, ya que puede almacenarse durante dos años y puede ser utilizado durante veinte meses por las autoridades participantes del proyecto de SyRI pertinente y, además, puede hacerse llegar a la Fiscalía o a la Policía, a petición de ellas. Ello implica que el propio informe de riesgo sobre una persona, aunque no conlleve ulterior tratamiento, dados sus efectos puede considerarse, de acuerdo con esta interpretación, como una decisión individual automatizada en los términos del artículo 22[531].

En uno de los casos Uber, el Tribunal de Distrito de Ámsterdam, al decidir sobre la aplicación del artículo 22 en relación con el despido de varios *drivers* o conductores, se detiene a argumentar si la suspensión temporal de la cuenta, previa a la decisión final, produce un efecto significativo puesto que, en tal caso, sería aplicable el artículo 22 no solo a la decisión final de despido, sino también a la suspensión temporal previa[532]. Por

[530] Kamarinou, Millard, y Singh, «Machine Learning with Personal Data», 12.

[531] Sentencia de la Corte de Distrito de La Haya [Rechtbank Den Haag] de 5 de febrero de 2020 (ECLI: NL: RBDHA:2020: 865), par. 6.59. El Tribunal sostuvo esta argumentación en base a las directrices del GT29, sosteniendo que dichos informes tenían el potencial de debe tener el potencial de: afectar significativamente a las circunstancias, al comportamiento o a las elecciones de las personas afectadas; tener un impacto prolongado o permanente en el interesado; o en los casos más extremos, provocar la exclusión o discriminación de personas -par. 6.36-, apoyado en Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 24.

[532] Sentencia del Tribunal de Distrito de Ámsterdam [Rechtbank Amsterdam] de 11 de marzo de 2021, *Uber deactivation case* (ECLI:NL:RBAMS:2021:1018), par 4.11. En este caso, siguiendo las Directrices del GT29, el Tribunal determinó que la suspensión temporal no producía tales efectos dado que no producía un impacto prolongado o permanente en el interesado, de acuerdo con Grupo de Trabajo sobre Protección de Datos del Artículo 29, 24. Ahora, de haber llegado a la conclusión contraria, esto es, de haber concluido que la suspensión temporal producía dichos efectos significativos, ello habría hecho valorar al Tribunal si había intervención humana previa a la suspensión temporal y, no siendo así -cuestión que sí resuelve en par. 4.25-, la suspensión temporal sería una decisión automatizada prohibida por el artículo 22(1), a pesar de ser una etapa intermedia o puntual en el procedimiento de despido.

ende, a la hora de evaluar los riesgos de los tratamientos automatizados que realiza para cumplir con el artículo 22 RGPD, el responsable debe atender, no tanto a la toma de decisiones final, sino a los efectos que puedan producirse en cada etapa que involucre este tipo de tratamiento automatizado, incluida la elaboración de perfiles[533].

Por otro lado, como consecuencia del enfoque basado en el riesgo, al responsable del tratamiento le corresponde determinar, en primer lugar, si dicho tratamiento produce o no estos efectos. Es decir, si el responsable lleva a cabo un tratamiento automatizado, incluida la elaboración de perfiles, deberá determinar en primer lugar si supera el umbral fijado por los efectos jurídicos o significativos, antes de evaluar, por ejemplo, si la decisión está o no basada únicamente en el tratamiento automatizado. Para ello, no obstante, es primordial que el responsable pueda disponer de unos criterios claros a la hora de realizar esta clase de evaluación, en definitiva, a la hora de responder a la pregunta, ¿qué debemos entender por efectos jurídicos o de afectación significativa similar?

Las directrices refrendadas por el CEPD aportan criterios en este sentido que ya han sido utilizados por distintos tribunales para determinar la producción o no de dichos efectos, tal y como hemos visto arriba en los casos SyRI y Uber, y también para el caso Ola[534]. Sin embargo, como veremos, estos criterios no son suficientes para aportar una seguridad jurídica aceptable en el ecosistema del RGPD[535].

En cuanto a la producción de efectos jurídicos, para las directrices del GT29 exige que la decisión afecte a los derechos o al estatuto jurídico de una persona, así como a los derechos adquiridos en virtud de un contrato[536]. Por su parte, la doctrina entiende que

[533] Esta evaluación puede resultar particularmente clave en entornos de regulación algorítmica, donde la actividad de los interesados está constantemente monitorizada. Vid. Apartado 3.2. Regulación algorítmica, en Introducción a la gobernanza y supervisión humana de la toma de decisiones automatizada basada en la elaboración de perfiles.

[534] En este caso el Tribunal de Distrito de Ámsterdam determinó que el sistema de "sanciones y deducciones" impuesto por Ola produce efectos significativos ya que afecta significativamente a las circunstancias, el comportamiento o las elecciones de los individuos afectados -par. 4.51-. Bajo mi punto de vista, también podrían considerarse efectos jurídicos, ya que afectan a los derechos contractuales adquiridos por la persona interesada. Sentencia del Tribunal de Distrito de Ámsterdam [Rechtbank Amsterdam] de 11 de marzo de 2021, *Ola transparency case* (ECLI:NL:RBAMS:2021:1019).

[535] Desde mi punto de vista, la ambigüedad del artículo 22 relativa a la determinación de los efectos jurídicos o significativos es una de las principales causas que ha llevado a ingresar en la UCI al artículo 22 RGPD.

[536] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 23. Como ejemplos cita que la decisión pueda provocar: *la cancelación de un contrato; el derecho o la denegación de una*

dichos efectos estarían limitados a los casos en que se modifica el estatus jurídico o se crean obligaciones legales para una persona[537]. Bayamlıoğlu desarrolla más dicha definición, sosteniendo que los efectos jurídicos pueden describirse como: «*todas las calificaciones establecidas por una norma jurídica, ya sea en forma de obligaciones, permisos, derechos, poderes; o en relación con la propia condición, como ciudadano/a, progenitor/a, cónyuge, deudor/a; o en relación con categorías de cosas, por ejemplo, bienes muebles, títulos negociables, o dominio público*»[538]. En un aspecto que ha recibido poca atención por parte de la doctrina, entiendo que esta definición es la más ajustada para la producción de efectos jurídicos.

Por otro lado, se ha considerado que determinar qué son efectos significativos resulta más complejo que determinar qué son efectos jurídicos[539]. Como bien señala la doctrina, una de las diferencias entre el artículo 15 DPD y el actual 22 RGPD es que aquél no determinaba un vínculo entre los efectos jurídicos y los efectos significativos, mientras que el RGPD añadió la coletilla "similares" a los efectos significativos[540]. Este es un punto de partida relevante que también recogen las directrices del GT29: «*el límite de importancia debe ser similar al de una decisión que produzca un efecto jurídico*». Es decir, el umbral de la significancia debe definirse a partir del umbral establecido para los efectos jurídicos[541].

Según Mendoza y Bygrave, es necesario que los efectos sean adversos para el interesado -aunque no lo sean del todo-, y cuanto más adversos sean, más probable es que se

prestación concedida por la ley, como la prestación por hijos o la ayuda a la vivienda; o la denegación de admisión en un país o la denegación de ciudadanía.

[537] Veale y Edwards, «Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling», 401. Para Hildebrandt tiene un sentido demasiado amplio, así, alega que en sentido estricto todo contrato tiene un efecto jurídico, incluido comprar el pan, en Hildebrandt, «Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning», 116.

[538] Bayamlıoğlu, «The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"», 4.

[539] Kamarinou, Millard, y Singh, «Machine Learning with Personal Data», 12.

[540] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 89; Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 69.

[541] Aunque este aspecto aporta algo de seguridad jurídica, no es suficiente dado el abanico de posibilidades en los ámbitos público y privado que encontramos para la toma de decisiones automatizada.

consideren significativos[542], si bien, el artículo 22(1) no distingue estrictamente entre efectos "negativos" y "positivos"[543]. El Considerando 71 aporta algunos ejemplos de lo que serían efectos significativos[544] y, a partir de los mismos, las Directrices del GT29 elabora una serie de criterios/ejemplos más elaborados [545], que resumo en la siguiente tabla:

Potencial grado de afectación a la persona interesada	**Categoría o ámbito de aplicación de la decisión**	**Circunstancias o características personales de la persona interesada**
Afectar significativamente a las circunstancias, al comportamiento o a las elecciones de las personas afectadas	Afectar a las circunstancias financieras	Vulnerabilidad financiera o socioeconómica
	Afectar al acceso a servicios sanitarios	Vulnerabilidad por minoría de edad

[542] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 88.

[543] Sí lo hace, por ejemplo, la ley belga de protección de datos al implementar el artículo 22(2)(b), si bien solo en relación a los efectos jurídicos, teniendo que resultar estos desfavorables, dice así: Art. 35. *Toute décision fondée exclusivement sur un traitement automatisé, y compris le profilage, qui produit des effets juridiques défavorables pour la personne concernée ou l'affecte de manière significative (...)* ("Loi relative àla protection des personnes physiques àl'égard des traitements de données àcaractère personnel"). Extraído de Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations», 12. También lo hace el análogo artículo 11 de la Directiva 680/2016, vid. Guzman Fluja, «Proceso penal y justicia automatizada».

[544] 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.

[545] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 24-25.Siguiendo estos criterios/ejemplos el Tribunal de Distrito de la Haya resolvió sobre varios parámetros que Uber y Ola utilizaban para regular la actividad de sus *drivers*, algunos de los cuales ya se han mencionado, el resto de menor relevancia pueden consultarse en Lazcoz, «Automated decision-making under Amsterdam's District Court judgements: Drivers v. Uber and Ola». A pesar del esfuerzo interpretativo realizado por el Tribunal en estos casos, hay una diferencia importante entre los ejemplos que se dan en las Directrices y los casos Uber y Ola. El WP29 describe sobre todo los casos en los que el interesado tiene relaciones ad hoc u ocasionales con el responsable del tratamiento. Por ejemplo, la solicitud de un crédito, la admisión a la universidad, la solicitud de la ciudadanía o el acceso al empleo. Como se ha dicho antes, los conductores de Uber y Ola están sujetos a entornos de regulación algorítmica, que son lo contrario de las relaciones ad hoc u ocasionales. Esta diferencia habría merecido un mayor esfuerzo interpretativo por parte del Tribunal de Distrito de Ámsterdam.

Tener un impacto prolongado o permanente en el interesado	Afectar o provocar una desventaja en el ámbito laboral	
Provocar la exclusión o discriminación de personas	Afectar al acceso a la educación	

Tabla 3. Elaborada por el autor a partir de las Directrices del GT29 (2018)

Una de las discusiones que más interés ha despertado en relación con la producción de efectos significativos es sobre si la publicidad personalizada en línea puede o no producir esta clase de efectos. Para el GT29 esta clase de publicidad no presenta, con carácter general, riesgos de carácter significativo. No obstante, añade, dependiendo de las características del caso sí podría llegar a presentar estos riesgos[546]. Mendoza y Bygrave entienden igualmente que con carácter general no puede decirse que la publicidad en línea provoque efectos significativos, en tanto las consecuencias significativas no pueden ser totalmente emocionales; si bien, cabe que se produzcan cuando se produce una discriminación que se traduce en términos consecuencias económicas[547].

Un paso más allá, Edwards y Veale consideran que estos fenómenos se están convirtiendo en algo destructivo para nuestra democracia[548], y se preguntan para quién debe ser significativa una decisión, ¿para el individuo en cuestión o para la sociedad en su conjunto?[549]. Desde un enfoque más positivista, Malgieri y Comandé centran su atención

[546] Cita así: *el nivel de intrusismo del proceso de elaboración de perfiles, incluido el seguimiento de las personas en diferentes sitios web, dispositivos y servicios; las expectativas y deseos de las personas afectadas; la forma en que se presenta el anuncio; o el uso de conocimientos sobre las vulnerabilidades de los interesados.* Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 24.

[547] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 89; Wachter, «Affinity profiling and discrimination by association in online behavioural advertising», 403.

[548] No olvidemos que la Comisión Europea, en su propuesta de Reglamento AIA, ha incluido entre los sistemas de IA prohibidos por representar un riesgo inadmisible, entre otros, artículo 5.1 AIA: *a) La introducción en el mercado, la puesta en servicio o la utilización de un sistema de IA que se sirva de técnicas subliminales que trasciendan la conciencia de una persona para alterar de manera sustancial su comportamiento de un modo que provoque o sea probable que provoque perjuicios físicos o psicológicos a esa persona o a otra; b) La introducción en el mercado, la puesta en servicio o la utilización de un sistema de IA que aproveche alguna de las vulnerabilidades de un grupo específico de personas debido a su edad o discapacidad física o mental para alterar de manera sustancial el comportamiento de una persona que pertenezca a dicho grupo de un modo que provoque o sea probable que provoque perjuicios físicos o psicológicos a esa persona o a otra. (…)*

[549] Edwards y Veale, «Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?», 48. Esta discusión apunta a cuestiones ya mencionadas planteadas a partir de los desarrollos tecnológicos más recientes, como la creación de grupos ad hoc a partir del uso de algoritmos de aprendizaje

en la Directiva 2005/29/CE sobre prácticas comerciales desleales[550], y concluyen que el tratamiento que pueda mermar de forma importante la libertad de elección o conducta del consumidor debe considerarse como un efecto significativo a efectos del artículo 22(1) RGPD[551].

Desde mi punto de vista, los criterios aportados por las Directrices del GT29 representan un intento loable de reducir la ambigüedad de la terminología utilizada en el artículo 22 por el legislador, sin embargo, se muestra insuficiente. El enfoque basado en el riesgo contenido en el artículo 22 RGPD exige que el responsable del tratamiento se responsabilice a la hora de determinar si los efectos producidos por el tratamiento de datos personales que realiza entran o no en el ámbito de aplicación de dicha disposición.

Ahora bien, el cumplimiento de tales obligaciones jurídicas – y su demostración – requieren de una seguridad jurídica incompatible con un artículo 22 ingresado en una unidad de cuidados intensivos. El CEPD debería contribuir con nuevas directrices abordando este aspecto de manera monográfica.

4. Reflexiones provisionales sobre el capítulo segundo – tentative thoughts on chapter two

En este apartado se recogen una serie de reflexiones provisionales a modo de cierre de cada capítulo. Aunque algunas de estas reflexiones servirán de apoyo para las conclusiones de esta investigación, el objetivo de este apartado no es exponer dichas conclusiones propiamente, sino resaltar de forma telegráfica algunos aspectos clave resultado del análisis realizado en cada capítulo.

---

automático susceptibles de ser discriminados, para las que el enfoque clásico de la privacidad, centrada en la atomización del individuo, puede mostrarse insuficiente.

[550] Directiva 2005/29/CE del Parlamento Europeo y del Consejo, de 11 de mayo de 2005, relativa a las prácticas comerciales desleales de las empresas en sus relaciones con los consumidores en el mercado interior, que modifica la Directiva 84/450/CEE del Consejo, las Directivas 97/7/CE, 98/27/CE y 2002/65/CE del Parlamento Europeo y del Consejo y el Reglamento (CE) n° 2006/2004 del Parlamento Europeo y del Consejo («Directiva sobre las prácticas comerciales desleales»)

[551] Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 253. Construyen dicha argumentación particularmente sobre la base del artículo 8 de la Directiva 2005/29/CE sobre prácticas comerciales agresivas: Se considerará agresiva toda práctica comercial que, en su contexto fáctico, teniendo en cuenta todas sus características y circunstancias, merme o pueda mermar de forma importante, mediante el acoso, la coacción, incluido el uso de la fuerza, o la influencia indebida, la libertad de elección o conducta del consumidor medio con respecto al producto y, por consiguiente, le haga o pueda hacerle tomar una decisión sobre una transacción que de otra forma no hubiera tomado.

- En la actualidad, la norma que regula la toma de decisiones automatizada de forma transversal a distintos contextos de aplicación es el RGPD, aunque solo lo hace directamente para la fase de implementación del ciclo de vida de los modelos algorítmicos. El artículo 22 adopta aquí un papel protagonista, aunque la regulación de la toma de decisiones automatizada por el RGPD debe analizarse en conjunto con el resto de derechos, obligaciones y principios aplicables.
- No sorprende encontrar esta regulación en el RGPD, puesto que la toma de decisiones automatizada ha encontrado acomodo históricamente en la normativa referida al desarrollo de los derechos a la vida privada y a la protección de datos, con las particularidades que los distintos cuerpos normativos tienen en cada contexto a la hora de adoptar y desarrollar estos derechos.
- En nuestro ordenamiento podemos observar un trazo más o menos reconocible entre los conceptos de intimidad, privacidad y protección de datos: la era informática obliga a la protección de una esfera personal más amplia del espacio tradicionalmente reservado a una intimidad que va incorporando una diversidad de intereses dignos de protección, esto es, la privacidad, y a su vez, el control sobre esa esfera personal se relaciona materialmente con la protección de datos.
- La posibilidad de generar bases de datos y su tratamiento informático alteraron la forma en que se toman decisiones y elaboran juicios que afectan a nuestras vidas, exacerbando y transformando los desequilibrios de poder existentes, lo cual provocó la necesidad de regular la elaboración de perfiles y la toma de decisiones automatizada en primera instancia.
- La introducción de complejos algoritmos de aprendizaje automático para el tratamiento de ingentes bases de datos con fines de perfilado individual y como base para la toma de decisiones de organizaciones públicas y privadas, ha supuesto un nuevo hito en esta problemática para la privacidad y la protección de datos personales.
- Al igual que su precedente en la Directiva de 1995, el artículo 22 RGPD es una disposición absolutamente kafkiana. Se encuentra repleta de ambigüedades y de excepciones a las prohibiciones que establece con carácter general. Su falta de claridad y simplicidad ha provocado que se considere un derecho de segunda

categoría, escasamente aplicado por los tribunales e interpretado de forma inconsistente por la doctrina.

- El derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado -22(1) RGPD- debe ser interpretado como una prohibición de carácter general para el responsable del tratamiento, y no como un derecho a interponer por la persona interesada en forma de objeción a dicha forma de tratamiento. Esta interpretación ha sido adoptada de forma mayoritaria por la doctrina, así como por las directrices del GT29 -refrendadas por el CEPD- y la jurisprudencia.
- La aplicación del artículo 22 depende de la producción de efectos jurídicos o de afectación significativa similar. Estos efectos establecen el umbral de riesgo mínimo necesario para la aplicación de las disposiciones sobre decisiones automatizadas, no obstante, es necesario contar con más y mejores criterios -también desde perspectivas sectoriales- para entender este umbral y que los responsables del tratamiento puedan operar con seguridad jurídica.

As a method of recapping each chapter, this section presents a number of tentative thoughts. The research's conclusions will be supported by some of these insights, but the purpose of this section is not to present those conclusions in their entirety. Rather, it aims to emphasize certain key aspects that came out of the analysis done in each chapter in a telegraphic manner.

- The GDPR currently regulates automated decision-making (ADM) across many application contexts, although it only does so directly for the deployment stage of the lifecycle of algorithmic models. While the regulation of automated decision-making under the GDPR must be examined in connection with all other applicable rights, obligations, and principles, Article 22 plays a significant role in this scenario.
- Given that automated decision-making has historically found a place in the laws pertaining to the protection of the rights to privacy and data protection -with the particularities that the various bodies of law have in each context when it comes to adopting and developing these rights-, it is not surprising to find the regulation of ADM in the GDPR today.

- In our legal system we can observe a more or less recognisable line between the concepts of privacy, intimacy and data protection: the computer era demands the protection of a personal sphere that is broader than the space traditionally reserved for intimacy, which is incorporating a diversity of interests worthy of protection, that is, privacy, and in turn, control over this personal sphere is materially related to data protection.
- The possibility of generating databases and their computer processing altered the way in which decisions and judgements affecting our lives are made, exacerbating and transforming existing power imbalances, which led to the need to regulate profiling and automated decision-making in the first place.
- The introduction of complex machine learning algorithms for the processing of huge databases for individual profiling purposes and as a basis for decision-making by public and private organisations has been a new milestone in this issue for privacy and personal data protection.
- Like its precedent in the 1995 Directive, Article 22 GDPR is an utterly Kafkaesque provision. It is full of ambiguities and exceptions to its general prohibitions. Its lack of clarity and simplicity has led to it being considered a second-class right, rarely applied by courts and inconsistently interpreted by the legal literature.
- The right not to be subject to a decision based solely on automated processing -22(1) GDPR- should be understood as a general prohibition for the controller, and not as a right to be invoked by the data subject in the form of an objection to that form of processing. This interpretation has been adopted by the majority of the legal literature, as well as by the WG29 guidelines -endorsed by the EDPB- and case law.
- The application of Article 22 depends on the production of legal effects or similar significant impact. These effects establish the minimum risk threshold necessary for the application of the provisions on automated decisions. However, more and better criteria -also from sectoral perspectives- are needed to understand this threshold so that controllers can operate with sufficient legal certainty.

CAPÍTULO 3. TRES PILARES SOBRE LOS QUE INTERPRETAR Y HACER EFECTIVA LA REGULACIÓN DE LA TOMA DE DECISIONES EN EL RGPD: DERECHO A LA INTERVENCIÓN HUMANA, DERECHO A LA INFORMACIÓN Y DERECHO A IMPUGNAR LA DECISIÓN. UNA PROPUESTA TERAPÉUTICA

CAPÍTULO 3. TRES PILARES SOBRE LOS QUE INTERPRETAR Y HACER EFECTIVA LA REGULACIÓN DE LA TOMA DE DECISIONES EN EL RGPD: DERECHO A LA INTERVENCIÓN HUMANA, DERECHO A LA INFORMACIÓN Y DERECHO A IMPUGNAR LA DECISIÓN. UNA PROPUESTA TERAPÉUTICA

Una vez realizado ese ejercicio diagnóstico sobre el artículo 22 y su contenido, este capítulo tiene por objetivo desviar el foco del análisis doctrinal que ha centrado el interés sobre esta disposición: los derechos de información y acceso y, en particular, la discusión relativa a la existencia o no de un derecho a una explicación para las decisiones automatizadas y su extensión[552]. Este enfoque no deja de ser relevante en este capítulo, pero pasará a ser uno de los tres pilares sobre los que reinterpretar la regulación de la toma de decisiones automatizada en el RGPD[553].

En este capítulo, se propone una interpretación que ofrezca vías útiles para la aplicación del artículo 22 o disposición kafkiana[554].

A través de los tres pilares que se proponen, pretendo poner de relieve que el conjunto de mecanismos de gobernanza, ex ante y ex post, que pone a disposición de la persona interesada el RGPD en la regulación de la toma de decisiones es más amplia de lo que muchos análisis han propuesto hasta la fecha. La perspectiva está centrada en los derechos del interesado, sin embargo, se pondrán de manifiesto varias de las limitaciones de establecer esta carga sobre los individuos.

¿Por qué la intervención humana, la información y la impugnación de la toma de decisiones automatizada? Esta idea surge al indagar sobre los trabajos preparatorios de la

[552] El esfuerzo por trascender este análisis responde también las limitaciones de los derechos de información y del enfoque del principio de transparencia en sí, al respecto Apartado 2. Derecho a la información en la toma de decisiones automatizada, en este mismo capítulo.

[553] Este enfoque pretende así devolver el protagonismo al RGPD en este contexto, dado que las limitaciones del anterior enfoque han provocado, en muchas ocasiones, que la atención doctrinal haya transitado hacia otros ámbitos del Derecho, por ejemplo, al derecho antidiscriminatorio que, a mi juicio, aunque útil en ciertos aspectos no dispone de la transversalidad que caracteriza al RGPD para la regulación de la toma de decisiones automatizada.

[554] Bayamlioglu resalta que, a pesar del flujo de artículos en la doctrina sobre transparencia, interpretabilidad o explicabilidad en torno a este artículo, hasta el momento se han ofrecido pocas vías prácticas para la efectiva aplicación de esta disposición, en Bayamlıoğlu, «The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"», 17. Esta investigación pretende ofrecer vías específicas para la aplicación de los mecanismos de intervención humana en el artículo 22 RGPD, sin dejar de lado otras vías abiertas por la doctrina.

Comisión Europea para la Directiva de Protección de Datos de 1995. Aunque se ahondará con profundidad en estos trabajos más adelante, hay dos temores fundamentales a la hora de regular esta clase de tratamiento de datos personales. Por un lado, hay un temor a la pérdida del control humano en las decisiones que uno mismo toma, como responsable del tratamiento, es decir un temor a la abdicación de la responsabilidad por aquél en favor de la máquina; por otro lado, hay un temor a esa pérdida de control humano en las decisiones que se toman sobre uno mismo, es decir, la pérdida de control sobre las decisiones que afectan a uno de forma significativa, como persona interesada.

Dichos temores expresan, a mi modo de ver, tres cuestiones que son fundamentales en la regulación de la toma de decisiones automatizada: la responsabilidad sobre el tratamiento [responsabilidad], la capacidad de la persona interesada para influir sobre el tratamiento [permeabilidad] y la posibilidad de observar que dicho tratamiento es permeable y responsable [transparencia].

En esta investigación, esas cuestiones se han vinculado respectivamente con distintos mecanismos de gobernanza del RGPD, ex ante y ex post, aunque ello no quiere decir que cada uno de estos mecanismos respondan únicamente a dichas cuestiones -puesto que están interconectadas-, sino que es su fundamento principal: la intervención humana como responsabilidad, el derecho a contestar la decisión como permeabilidad y los derechos de información como transparencia.

Los tres pilares se edifican de la siguiente forma[555]. La intervención humana debe asegurar una supervisión apropiada y significativa de la toma de decisiones automatizada sobre la base del principio de responsabilidad que atañe al responsable del tratamiento. Los derechos de información deben permitir que la persona interesada adquiera un conocimiento suficiente sobre el tratamiento automatizado de sus datos personales que deriven en inferencias sobre su persona o en la adopción de decisiones que le afecten de forma significativa. Por último, el derecho a impugnar la decisión ha de permitir a la persona interesada rebatir y contestar las inferencias y decisiones adoptadas sobre la base de sus datos personales, asegurando el control sobre los mismos y las consecuencias de

[555] El grueso de la investigación se centra en la intervención humana, primero, por haber sido uno de los aspectos menos trabajados por la doctrina y, segundo, porque el conjunto de esta investigación tiene por objeto centrarse en la supervisión de los sistemas automatizados. Ahora bien, ello no quiere decir que este "pilar" tenga más importancia que el resto en el enfoque conjunto.

su tratamiento y el respeto a sus derechos y libertades. Por supuesto, bajo este enfoque el artículo 22 no actúa de forma autónoma sino dentro del ecosistema del RGPD, lo que hace necesario el acercamiento al resto de derechos, obligaciones y principios del mismo.

En definitiva, el enfoque aportado a continuación pretende que el artículo 22 deje de considerarse un derecho de segunda clase y que -junto al resto de garantías del RGPD- pueda convertirse en lo que Hildebrandt considera que debería ser la regulación de estos sistemas bajo la normativa de protección de datos; una contribución fundamental para restablecer los controles y equilibrios en la relación entre nosotros, los humanos, y las máquinas que reconfiguran nuestro entorno, permitiendo impugnar la exactitud, la relevancia y la fiabilidad de estos sistemas[556].

1. Derecho a la intervención humana

Si observamos cómo se ha analizado la intervención humana en el artículo 22, vemos que habitualmente solo se habla de la intervención humana como medida de salvaguarda en el apartado 22(3), e incluso se afirma que esta disposición no es aplicable a sistemas de apoyo a la toma de decisiones[557]. Esta interpretación debe ser objeto de crítica ya que limita los mecanismos de gobernanza basados en la intervención humana contenidos en el artículo 22. Estos mecanismos, ya identificados perfectamente en las directrices refrendadas por el CEPD, adquieren además una importancia destacada a la luz de las propuestas de regulación de IA, en relación con la supervisión humana como requisito de obligado cumplimiento[558].

Los siguientes apartados se desarrollan de la siguiente forma. Primero, sobre el enfoque basado en el riesgo descrito anteriormente, se diferencian tres tipos de decisiones en torno a este artículo (decisiones fuera del ámbito de aplicación del artículo 22 basadas o no únicamente en el tratamiento automatizado; decisiones no basadas únicamente en el tratamiento automatizado; decisiones basadas únicamente en el tratamiento

[556] Hildebrandt, «Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning», 118.

[557] Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 66.

[558] Siguiendo aquí lo desarrollado en el marco teórico sobre la supervisión humana y los sistemas automatizados. Apartado 2. La intervención humana en la toma de decisiones automatizada basada en la elaboración de perfiles, en Capítulo 1. Marco teórico de la toma de decisiones automatizada basada en la elaboración de perfiles.

automatizado) y, a partir de aquí, se describe cómo los distintos mecanismos de intervención humana contenidos en el apartado primero (*human in the loop*) y en el apartado tercero (*human out of the loop*) se integran respectivamente en las decisiones no basadas únicamente en el tratamiento automatizado y en las basadas únicamente en el mismo. Segundo, se explica detalladamente que el RGPD, a la luz de las directrices refrendadas por el CEPD, requiere de una intervención humana significativa. Ahora bien, definir qué debe entenderse por intervención humana significativa no queda tan claro en dichas directrices. Por ello, a continuación, se indaga en el fundamento de la regulación del artículo 22 y su predecesor -art.15 DPD- para tratar de desarrollar el concepto significativo para la intervención humana. Aquí, se destaca la conexión entre la intervención humana y la responsabilidad sobre el cumplimiento y su demostración, sobre la que se ahondará en el siguiente capítulo.

1.1. Dos mecanismos de intervención humana distintos en los tres tipos de decisiones automatizadas en torno al artículo 22.

Efectivamente, en la doctrina es habitual ver análisis de esta disposición que se refieren exclusivamente a la intervención humana en el apartado tercero de la misma, donde se hace explícita entre las medidas de salvaguarda para los derechos y libertades de las personas interesadas ante la toma de decisiones basada únicamente en el tratamiento automatizado. Así, se obvia que la prohibición de tomar decisiones basadas únicamente en el tratamiento automatizado obliga, necesariamente, a definir qué se entiende por "basadas únicamente". Como se ha señalado anteriormente, éste es uno de los aspectos más ambiguos de ese apartado primero. Y es donde se introduce el primero de los mecanismos de intervención humana, donde el RGPD establece una clara preferencia político-jurídica por los sistemas *in the loop*[559].

Antes de entrar en ello, volvamos sobre el enfoque basado en el riesgo definido también anteriormente y veamos qué clases de decisiones podemos distinguir en torno al artículo 22.

Por un lado, tenemos decisiones que pueden o no producir un efecto jurídico o de afectación significativa similar. Por otro lado, las decisiones pueden o no estar basadas

[559] Hoofnagle, van der Sloot, y Borgesius, «The European Union general data protection regulation: what it is and what it means», 68.

únicamente en el tratamiento automatizado. Sin embargo, si una decisión no produce ese efecto jurídico o significativo, es decir, si se coloca por debajo del umbral de ese enfoque basado en el riesgo, será indiferente que la misma esté o no basada únicamente en el tratamiento automatizado.

Ello quiere decir que la primera clase de decisiones que podemos definir son [1] las decisiones que no producen efectos jurídicos o significativos estén o no basadas únicamente en el tratamiento automatizado. Ahora bien, entre las decisiones que sí producen dichos efectos, debemos distinguir necesariamente dos clases de decisiones; [2] las decisiones que producen efectos jurídicos o significativos que no están basadas únicamente en el tratamiento automatizado – legítimas conforme al apartado 22(1) –, esto es, sistemas de apoyo a la toma de decisiones; y [3] las decisiones que producen efectos jurídicos o significativos y que están basadas únicamente en el tratamiento automatizado – prohibidas con carácter general por el apartado 22(1) –.

Las decisiones que no superan el umbral del riesgo [1], no entran en el ámbito de aplicación del artículo 22. En cambio, sí entran en su ámbito de aplicación las decisiones que alcanzan ese umbral de riesgo [2] y [3] y, para cada una, el RGPD establece un mecanismo obligatorio de intervención humana distinto.

Desde esta perspectiva, siempre que se produce un efecto jurídico o significativo, la prohibición del apartado primero introduce la intervención humana como componente esencial de la toma de decisiones automatizada [2], en forma de *human in the loop*. Así, el responsable del tratamiento puede esquivar la prohibición siempre que no base la decisión únicamente en el tratamiento automatizado, es decir, introduciendo un agente humano en el proceso de toma de decisiones.

Por otro lado, también cabe la toma de decisiones basada únicamente en el tratamiento automatizado cuando se acuda a alguna de las excepciones contenidas en el apartado segundo, es decir, a las excepciones para la prohibición general contenida en el apartado primero[560]. En este caso, el RGPD legitima la toma de decisiones basada únicamente en el tratamiento automatizado que produce efectos jurídicos o significativos [3], ahora bien, condiciona dicha legitimación al cumplimiento de las medidas de salvaguarda contenidas

[560] Si la decisión se adopta sobre categorías especiales de datos a las que se refiere el artículo 9 RGPD, las excepciones están limitadas a las establecidas por el apartado 4 del artículo 22 RGPD.

en el apartado tercero[561], entre las cuales, el RGPD introduce la intervención humana como medida de salvaguarda bajo requerimiento del interesado, en forma de *human out of the loop*.

Ello implica que los distintos mecanismos de intervención humana se integran en tipos de decisiones diferentes y, por ende, en distintas fases de la toma de decisiones; en el primer caso, la intervención humana como componente esencial de la toma de decisiones automatizada tiene lugar antes de la producción de efectos jurídicos o significativos para la persona interesada; mientras que la intervención humana como medida de salvaguarda bajo requerimiento del interesado tiene lugar tras la producción de dichos efectos, tal y como puede observarse en la siguiente tabla[562]:

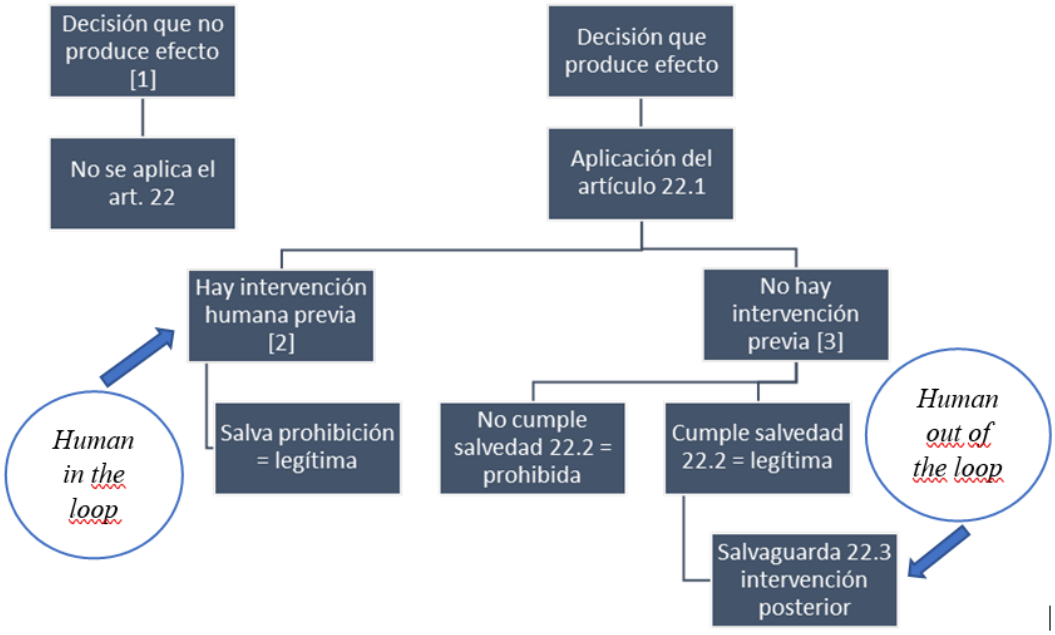

Tabla 4. Elaborada por el autor

[561] Salvo, como ya hemos visto, para el caso de que la decisión automatizada se encuentre autorizada por el Derecho de la Unión o de los EEMM, en cuyo caso las medidas de salvaguarda no son necesariamente las contenidas en el apartado tercero; en cualquier caso, la norma debe establecer medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado.

[562] No obstante, en lo relativo a la salvaguarda del 22(3), ha de recordarse que admite excepciones tanto cuando se aplica la salvedad contenida en 22(2)(b), es decir, habilitado por el Derecho de la Unión o de los EEMM, como cuando el tratamiento está basado en categorías especiales de datos personales y, por ende, han de aplicarse las salvedades contenidas en 22(4). Vid. Apartado 3.2. Dos prohibiciones y un sinfín de excepciones, en Capítulo 2. Toma de decisiones automatizada basada en la elaboración de perfiles en el RGPD: el artículo 22 en la unidad de cuidados intensivos.

Observemos ahora esta distinción con mayor detenimiento, en este sentido nos será útil acudir a los ejemplos de mecanismos de gobernanza basados en la intervención humana aportados por el Libro Blanco de IA[563].

1.1.1. Intervención humana como componente esencial de la toma de decisiones automatizada en el RGPD: *human in the loop*

Anteriormente en esta investigación, hemos visto las diferencias entre la intervención humana en forma de *human in the loop*, *human on the loop* o *human out of the loop*. En particular, en el Libro Blanco de IA publicado por la Comisión Europea en febrero de 2020 se ilustraban las diferencias entre estos mecanismos, ejemplificando que el tipo y nivel adecuado de supervisión humana puede variar de un caso a otro.

El primero de esos mecanismos se define así: «*El resultado del sistema de IA no es efectivo hasta que un humano no lo haya revisado y validado (por ejemplo, la decisión de denegar una solicitud de prestaciones de seguridad social solo podrá adoptarla un ser humano)*»[564]. Como ya habíamos adelantado, aquí el Libro Blanco de IA está recogiendo el mecanismo introducido por el artículo 22 en su apartado primero al prohibir la toma de decisiones automatizada. Las decisiones basadas únicamente en el tratamiento automatizado que produzcan un efecto jurídico o significativo están prohibidas con carácter general por este apartado, lo cual significa que serán legitimas aquellas decisiones que incorporen intervención humana a la toma de decisiones basada en el tratamiento automatizado de datos, incluida la elaboración de perfiles, incorporando dicha intervención *in the loop*.

En otras palabras, una decisión automatizada que genere un efecto jurídico o significativo debe contar con intervención humana previa a la producción de dicho efecto para que ésta

[563] Para establecer esta distinción, es importante señalar que, en el caso del Libro Blanco, la distinción radica en si la intervención humana se garantiza antes o después de que el resultado sea "efectivo", mientras que en el RGPD radica en si la intervención humana se garantiza antes o después de que el resultado produzca un efecto jurídico sobre el interesado o le afecte significativamente de modo similar.

[564] Comisión Europea, Libro Blanco sobre la inteligencia artificial, 25.

sea legítima conforme al 22(1)[565]. Tal y como señala Jones, la prohibición general se traduce en un derecho al *human in the loop*[566].

La distinción entre los mecanismos de intervención humana en el RGPD para las decisiones que producen efectos jurídicos o significativos, [2] y [3] respectivamente, no responde únicamente a la fase de la toma de decisiones en que tienen lugar, antes o después de la producción de dichos efectos, sino también al distinto objetivo normativo o fundamento al que responde cada clase de intervención[567]. En este sentido, la intervención humana introducida por la prohibición del apartado primero tiene por objeto garantizar el derecho a no ser objeto de decisiones individuales automatizadas basadas únicamente en el tratamiento automatizado. Es decir, el RGPD introduce aquí la intervención humana como un componente esencial de la toma de decisiones individuales automatizadas, incluida la elaboración de perfiles, para las decisiones no basadas únicamente en el tratamiento automatizado [2][568].

Esta idea parece cristalizarse en la redacción del Considerando 71 y su primera mención a la intervención humana, cuando dice: «*El interesado debe tener derecho a no ser objeto de una decisión, que puede incluir una medida, que evalúe aspectos personales relativos a él, y que se base únicamente en el tratamiento automatizado y produzca efectos jurídicos en él o le afecte significativamente de modo similar, como 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*».

Por un lado, la prohibición pretende que los interesados no sean objeto de decisiones que produzcan determinados efectos sin intervención humana alguna; por otro, se constituye la introducción de la intervención humana como una garantía normativa para el derecho

[565] Salvo que dicha decisión basada únicamente en el tratamiento automatizado esté legitimada por alguna de las excepciones contenidas en el apartado 22(2), en cuyo caso, el responsable del tratamiento podrá esquivar la introducción de intervención humana como un componente esencial de la toma de decisiones. No obstante, deberá facilitar la intervención humana como medida de salvaguarda conforme al apartado 22(3) en los términos que veremos a continuación.

[566] Jones, «The right to a human in the loop: Political constructions of computer automation and personhood», 224.

[567] Sobre el fundamento de la intervención humana en el RGPD profundizaremos más adelante. Vid. Apartado 1.2.2. Fundamento para la intervención humana en el artículo 22 RGPD, en este mismo capítulo.

[568] Acerca de cómo ha de ser esta intervención, en términos cualitativos, también entraremos más adelante. Vid. Apartado 1.2. Intervención humana significativa: la necesidad de superar un concepto formal de intervención humana, en este mismo capítulo.

a no ser objeto de esta clase de decisiones prohibidas. De este modo, el RGPD asegura que la toma de decisiones incluye también un razonamiento humano y, por ende, el tratamiento automatizado no es el fundamento exclusivo -basadas únicamente- de la toma de decisiones[569].

1.1.2. Intervención humana como medida de salvaguarda bajo requerimiento del interesado: *human out of the loop*

Volvamos a las distintas manifestaciones de la intervención humana como forma de garantizar una adecuada supervisión de los sistemas automatizados que recoge el Libro Blanco de IA. En segundo lugar, recoge: «*El resultado del sistema de IA es inmediatamente efectivo, pero se garantiza la intervención humana posterior (por ejemplo, la decisión de denegar una solicitud de tarjeta de crédito puede tramitarse a través de un sistema de IA, pero debe posibilitarse un examen humano posterior)*». Este es el caso de la medida de salvaguarda contenida para la toma de decisiones totalmente automatizada en el 22(3) para las decisiones basadas únicamente en el tratamiento automatizado [3], el efecto jurídico o significativo se produce con anterioridad a la intervención humana.

Recordemos que, en este caso, la toma de decisiones basada únicamente en el tratamiento automatizado está excepcionalmente legitimada por las cláusulas recogidas en el apartado 22(2), si bien, dichas excepciones están condicionadas a la adopción de medidas de salvaguarda contenidas en el apartado 22(3)[570], entre las cuales, encontramos el derecho a obtener intervención humana por parte del responsable para las decisiones basadas únicamente en el tratamiento automatizado [3].

En este caso, el efecto jurídico o significativo tiene lugar de forma previa al mecanismo de intervención humana dado que se legitima excepcionalmente bajo alguna de las cláusulas legitimadoras contenidas en dicho apartado segundo. No obstante, una vez más el RGPD introduce la intervención humana con carácter obligatorio, de modo que esta vía excepcional -para la toma de decisiones basadas únicamente en el tratamiento

[569] Wagner, «Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems», 108.

[570] A continuación, entraremos en las matizaciones que admite esta regla general.

automatizado- solo es legítima en tanto se garantice el derecho a la intervención humana posterior, un derecho al *human out of the loop*, entre otras medidas de salvaguarda.

Así, el Considerando 71 menciona por segunda la intervención humana: «*Sin embargo, se deben permitir las decisiones basadas en tal tratamiento, (…). En cualquier caso, dicho tratamiento debe estar sujeto a las garantías apropiadas, entre las que se deben incluir la información específica al interesado y el 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*».

El RGPD, no obstante, coloca este segundo mecanismo de intervención humana como una medida a disposición de la persona interesada, es decir, a ejercer o interponer a discreción de la misma[571]. Para Almada, este derecho a obtener intervención humana tiene como fundamento la reevaluación de una determinada forma de tratamiento automatizado[572], generalmente prohibida y solo excepcionalmente autorizada. Tanto el fundamento normativo de este mecanismo de intervención humana, como el hecho de que se establezca como una medida bajo requerimiento del interesado, ilustran que este segundo mecanismo –*human out of the loop*– tiene una importancia menor en el ecosistema del RGPD.

Tal y como se ha adelantado, la aplicación de la intervención humana como medida de salvaguarda para los derechos y libertades de las personas interesadas ha de matizarse en dos supuestos. Por un lado, cuando esta clase de decisiones [3] se habilita por la cláusula 22(2)(b), es decir, se autoriza por el Derecho de la Unión o de los EEMM, no corresponde -directamente- la aplicación de las medidas de salvaguarda contenidas en el 22(3), dicha cláusula hace simplemente una referencia genérica al establecimiento por dichas normas

[571] El hecho de que el apartado primero constituya una prohibición general, tal y como se ha expuesto anteriormente en esta investigación, implica que la intervención humana como componente esencial de la toma de decisiones debe asegurarse por el responsable del tratamiento con independencia de si el interesado lo requiere o no. En palabras de Binns, introducir de esta forma la intervención humana reduce la posibilidad de asegurar la justicia individual para los interesados: *(…), it means that decision-makers are likely to only review false positives (where people have been incorrectly been denied a benefit), and ignore false negatives (where people have incorrectly been granted a benefit), because the latter have no incentive to challenge a positive decision*. Binns, «Human Judgment in algorithmic loops: Individual justice and automated decision-making», 11.

[572] Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems», 1.De acuerdo con la interpretación realizada por Bayamlıoğlu, en esta investigación considero que la intervención humana como medida de salvaguarda es instrumental a la verdadera columna vertebral de las medidas del 22(3), esto es, el derecho a impugnar la decisión. Bayamlıoğlu, «The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"», 5.

de *medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado*[573]. Por otro lado, cuando el tratamiento automatizado se basa en las en las categorías especiales de datos personales contempladas en el artículo 9 RGPD, ha de aplicarse la segunda prohibición recogida en el apartado 22(4), lo cual implica que las excepciones a la misma también difieren, e igualmente se hace referencia genérica a que se tomen medidas adecuadas para la salvaguarda de derechos y libertades, no conectando dicha referencia con el apartado 22(3)[574].

1.2. Intervención humana significativa: la necesidad de superar un concepto formal de intervención humana

Una vez definidos los distintos mecanismos de intervención humana contenidos por el artículo 22 RGPD, debe determinarse el aspecto cualitativo de éstos, esto es, qué clase de intervención se requiere por parte de esta norma. En este apartado, desarrollaré -conforme a las directrices refrendadas por el CEPD- que la clase de intervención requerida por el RGPD es una intervención humana significativa para ambos mecanismos. No obstante, dado que la intervención humana como componente esencial de la toma de decisiones automatizada tiene un lugar destacado en el Reglamento -22(1)-, centraremos la argumentación en este apartado sobre el mismo, aunque también se realizarán por supuesto las referencias necesarias a la intervención humana como medida de salvaguarda bajo requerimiento del interesado.

[573] En las distintas legislaciones de los EEMM que han abordado la regulación de las decisiones automatizadas a partir del 22(2)(b) han recogido también como medida de salvaguarda el derecho a obtener intervención humana, según Malgieri al menos Bélgica, Países Bajos, Alemania, Irlanda, Hungría y también indirectamente Reino Unido en la Data Protection Act 2018, en Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations», 22. Si bien, ello no quiere decir que necesariamente se deba recoger la misma entre las medidas de salvaguarda.

[574] Con carácter general, a pesar de que el GT29 no aborda esta cuestión en sus directrices, podría decirse que las medidas de salvaguarda para excepciones contenidas en el apartado 22(4) deberán establecer, como mínimo, el nivel de protección asegurado por las medidas de salvaguarda para las excepciones del 22(2), dado que la prohibición del 22(4) es una prohibición agravada respecto del 22(1). Para el tratamiento legitimado por el artículo 9(2)(g), por razones de interés público esencial sobre la base del Derecho de la Unión o de los EEMM, tendremos que acudir a dicha normativa para observar si se incluye un derecho a la intervención humana como medida de salvaguarda, como hemos visto para la excepción del 22(2)(b). En cuanto al tratamiento legitimado por el artículo 9(2)(a), por consentimiento explícito del interesado, cabe interpretar de forma analógica que deben aplicarse por el responsable del tratamiento, al menos, las medidas establecidas para la excepción del 22(2)(c) contenidas en el 22(3), con lo cual, siempre incluiría el derecho a la intervención humana por parte del responsable.

Esta tarea requiere determinar el umbral mínimo de intervención humana requerido por el RGPD para considerar que una decisión no se basa únicamente en el tratamiento automatizado[575], si bien, la vaguedad del término en sí dificulta este análisis de la intervención humana[576].

Una parte de la doctrina, al analizar la aplicación de los derechos de información y acceso para las decisiones automatizadas, ha criticado que la mera inclusión formal de intervención humana es suficiente para considerar que una decisión no está basada únicamente en el tratamiento automatizado y excluye, por ende, la aplicación de los derechos de información y acceso de los artículos 13(2)(f), 14(2)(g) and 15(1)(h) RGPD[577]. Sobre dicha exclusión entraremos más adelante, lo determinante por el momento es considerar si esa inclusión "formal" de la intervención humana[578] constituye o no una decisión basada únicamente en el tratamiento automatizado.

A pesar de la mencionada crítica, este punto parece reunir cierto consenso. Las directrices del GT29 son claras al establecer que la intervención humana no puede ser únicamente un gesto simbólico[579]. Siguiendo esta misma postura encontramos voces autorizadas en la doctrina. Para Brkan una interpretación formalista que involucre al ser humano como parte necesaria del proceso de toma de decisiones, pero que en última instancia deje el poder de decisión a la máquina, no garantizaría un nivel suficientemente alto de protección de datos exigido por el RGPD[580]. Malgieri y Comandé apuntan a que cualquier

[575] Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 244.

[576] Gil González y Hert, «Understanding the legal provisions that allow processing and profiling of personal data—an analysis of GDPR provisions and principles», 617.

[577] Wachter, Mittelstadt, y Floridi, «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation», 88. Esta interpretación se apoya sobre la sentencia de 28 de enero de 2014 del Tribunal Federal de Justicia de Alemania o *Bundesgerichtshof* (BGH) para el caso SCHUFA sobre valoraciones crediticias en aplicación del artículo 15 DPD - *Scoring und Datenschutz*. 28. 1. 2014-VI ZR 156/13-.

[578] Para Roig este es un caso problemático en tanto la decisión "humana" se base en exclusiva en fuentes automatizadas, y también en tanto los sistemas incorporen algún humano en el proceso, pero no le faculten con la capacidad real de alterar la decisión automatizada, en Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 67. Veale y Edwards hacen referencia al fenómeno del *rubber stamping* o estampado de sellos, en Veale y Edwards, «Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling», 400.

[579] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 23.

[580] Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 101.

decisión humana de carácter pasivo entra en el ámbito de la prohibición del 22(1) como decisión basada únicamente en el tratamiento automatizado[581]. Para Mendoza y Bygrave, aunque una decisión se atribuya formalmente a una persona, se considerará que se basa únicamente en un tratamiento automatizado si una persona no evalúa activamente el resultado del tratamiento antes de su formalización como decisión[582].

Sobre este consenso, con base en el cual una intervención humana formal o pasiva debe considerarse equivalente a una decisión basada únicamente en el tratamiento automatizado; *a contrario sensu*, podría llegarse a la conclusión de que un proceso de toma de decisiones en el que un humano participa de forma activa y tiene una influencia real sobre las decisiones finales, no sería considerado como una decisión basada únicamente en el tratamiento automatizado. Ahora bien, ¿cómo puede definirse esta clase de intervención humana en el marco del RGPD?

Aquí, las directrices refrendadas por el CEPD establecen un concepto clave: intervención humana *significativa*.

Dicho término se introduce, en primer lugar, al definir cómo las decisiones automatizadas pueden solaparse parcialmente con la elaboración de perfiles, se aporta el siguiente ejemplo en el cual una decisión no basada únicamente en el tratamiento automatizado incluye una elaboración de perfiles: «*(…), antes de conceder una hipoteca, un banco puede tener en cuenta la calificación crediticia del prestatario, y pueden producirse otras intervenciones humanas significativas adicionales antes de que se tome ninguna decisión sobre la persona*»[583]. Es decir, las decisiones no basadas únicamente en el tratamiento se relacionan con la introducción de intervenciones humanas significativas.

[581] Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 252.

[582] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 4. La agencia de protección de datos de Reino Unido, Information Commissioner's Office (ICO), recoge en sus directrices sobre decisiones automatizadas y elaboración de perfiles que la participación humana debe tener un carácter activo y no meramente simbólico: *The question is whether a human reviews the decision before it is applied and has discretion to alter it, or whether they are simply applying the decision taken by the automated system*. ICO, «Guide to the UK General Data Protection Regulation (UK GDPR)», 8.

[583] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 9.

De nuevo, entre las formas de utilizar la elaboración de perfiles, elabora la distinción entre la toma de decisiones basada en la elaboración de perfiles y la toma de decisiones basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles[584], según las directrices esta última se aprueba y traslada a la persona interesada sin ninguna evaluación previa y significativa por parte de un ser humano[585]. De forma que las directrices reafirman que dicha intervención significativa debe tener lugar antes de la producción de efectos jurídicos o significativos, en forma de *human in the loop*.

Por último, a la hora de analizar qué clase de participación humana evita la toma de decisiones basadas únicamente en el tratamiento automatizado, las directrices ponen el foco sobre las obligaciones del responsable. El responsable del tratamiento no puede eludir la prohibición "fabricando" una intervención humana y, por lo tanto, los responsables del tratamiento deben garantizar que cualquier intervención humana sea significativa para el proceso de toma de decisiones del apartado 22(1)[586], y lo mismo establece para la intervención humana que debe proporcionarse como medida de salvaguarda de las decisiones totalmente automatizadas, habilitadas excepcionalmente por el apartado 22(2)[587].

La intervención humana significativa ha sido recientemente recogida por la jurisprudencia neerlandesa a la que se ha aludido más arriba. El Tribunal de Distrito de Ámsterdam en los tres casos relativos a Uber y Ola interpretó la prohibición del artículo 22 estableciendo que: «*una decisión basada únicamente en el tratamiento automatizado*

[584] En los términos expuestos anteriormente, pero poniendo el foco sobre la elaboración de perfiles, esta sería la misma diferenciación realizada para las [2] las decisiones que producen efectos jurídicos o significativos que no están basadas únicamente en el tratamiento automatizado – legítimas conforme al apartado 22(1) –; y [3] las decisiones que producen efectos jurídicos o significativos y que están basadas únicamente en el tratamiento automatizado – prohibidas con carácter general por el apartado 22(1) –.

[585] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 9.

[586] Grupo de Trabajo sobre Protección de Datos del Artículo 29, 23. Es cierto que el responsable del tratamiento puede eludir a través de la intervención humana los derechos de información y acceso establecidos para las decisiones basadas únicamente en el tratamiento automatizado. Ahora, el RGPD no da una carta blanca al responsable del tratamiento, transformar una decisión basada únicamente en el tratamiento en una decisión basada en la elaboración de perfiles requiere aumentar *significativamente el nivel de intervención humana de forma que el modelo ya no sea un proceso de toma de decisiones totalmente automatizadas*. Grupo de Trabajo sobre Protección de Datos del Artículo 29, 33. Las dimensiones de esta problemática se abordarán más adelante.

[587] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 30.

tiene lugar cuando no hay una intervención humana significativa en el proceso de toma de decisiones»[588].

Este concepto ha sido igualmente recogido por la AEPD en su Guía sobre la adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial: «*Para que pueda considerarse que existe participación humana, la supervisión de la decisión ha de ser realizada por una persona competente y autorizada para modificar la decisión, y para ello ha de realizar una acción significativa y no simbólica*»[589].

Por tanto, podría concluirse que la clase de intervención humana requerida bajo el RGPD es significativa. Sin embargo, es de lamentar que el artículo 22 no incluyese una mención expresa a este aspecto cualitativo de la intervención humana, al igual que debería haberse incluido explícitamente la intervención humana en el apartado 22(1), tal y como se hace para el apartado 22(3)[590].

1.2.1. Delimitación del concepto significativo

Ahora bien, determinar qué debemos entender por intervención humana significativa no es sencillo. En el anterior apartado, sobre el análisis doctrinal hemos podido adelantar, por un lado, que una intervención humana formal o pasiva no puede considerarse significativa y, por otro, que la participación humana debe ser activa y tener una influencia real sobre las decisiones finales. No obstante, la doctrina ha señalado las enormes dificultades prácticas a la hora de determinar esa gama de grises intermedia en

[588] See paragraph 4.63. *Uber transparency request* case (C/13/687315 / HA RK 20-207), paragraph 4.37. *Ola transparency request* case (C/13/689705 / HA RK 20-258) and paragraph 4.10. U*ber deactivation* case (C/13/692003 / HA RK 20-302) [*Van een uitsluitend op geautomatiseerde verwerking gebaseerd besluit is sprake indien er geen betekenisvolle menselijke tussenkomst is in het besluitvormingsproces]*.

[589] Agencia Española de Protección de Datos (AEPD), «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción», 10.

[590] En términos similares en los que Noto La Diega reclamaba un derecho a no ser objeto de decisiones en las que un humano no tome la decisión final, en sus propias palabras: *Therefore, it would seem more appropriate to recognise the right not to be subject to an algorithmic decision every time that there is not a human being clearly taking the final decisio*n. Noto La Diega, «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information», 19.

la toma de decisiones automatizada[591], lo cual no obsta para que esta labor jurídica sea muy necesaria[592].

En el capítulo anterior hemos tenido la oportunidad de hacer referencia al concepto de Control Humano Significativo (CHS) para Sistemas de Armas Autónomos Letales (SAAL), habiendo constatado algunas de las dificultades a las que el Derecho Internacional Humanitario (DIH) se ha enfrentado a la hora de delimitar dicho concepto[593]. También hemos constatado que en las distintas propuestas para la regulación de la IA en el ámbito normativo europeo se manejan conceptos similares en relación con la supervisión humana como requisito obligatorio para el desarrollo y uso de estos sistemas[594].

Volviendo al marco del RGPD[595], una vez más, en las directrices del GT29 encontramos un primer esfuerzo por delimitar este concepto, aportando algunos elementos comunes para entender la intervención humana significativa, tanto como un componente esencial de la toma de decisiones, *human in the loop*, como en forma de medida de salvaguarda para decisiones basadas únicamente en el tratamiento automatizado, *human out of the loop*.

---

[591] Edwards y Veale, *Slave to the Algorithm? Why a Right to Explanationn is Probably Not the Remedy You are Looking for*, 18:46-48; Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 89 y 93; Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 108.

[592] Almada destaca la importancia de identificar si la intervención humana en el proceso de toma de decisiones es significativa o meramente nominal: *In a scenario where the lines between full and partial automation are blurred, individuals might find themselves uncertain of the adequate channels for recourse, and this lack of information may cause delays or even block the reparation of harms caused by automation*. Almada, «Automated Decision-Making as a Data Protection Issue», 7.

[593] Sobre este concepto la doctrina ha destacado la necesidad de extender la aplicación del CHS a otros ámbitos del Derecho en relación con el uso de sistemas de IA y la relevancia del término "significativo" en esta jurídica. Vid. Romeo Casabona, «Criminal responsability of robots and autonomous artificial intelligent systems?», 183.

[594] El Libro Blanco sobre la IA de la Comisión establece que la supervisión humana debe alcanzarse garantizando una participación *adecuada* de las personas, en Comisión Europea, Libro Blanco sobre la inteligencia artificial, 25. La propuesta de Reglamento sobre los principios éticos de la IA del Parlamento hace referencia a que las decisiones adoptadas deben ser objeto de revisión, evaluación, intervención y control humanos *significativos* (Considerando 10), en 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)). Por último, la propuesta de Reglamento AIA, la Comisión establece la obligación de diseñar y desarrollar sistemas que puedan ser supervisados *de forma efectiva* por personas físicas al hacer uso de ellos -art. 14.1 AIA-.

[595] Lamentablemente, tal y como señala Koivisto, a pesar de que el RGPD sí utiliza el término significativo de forma explícita para los derechos de información y acceso, tampoco se define el mismo para su aplicación conforme a los artículos 13(2)(f), 14(2)(g) y 15(1)(h) RGPD, en Koivisto, «Thinking Inside the Box: The Promise and Boundaries of Transparency in Automated Decision-Making», 17.

En primer lugar, enlazando con esa pasividad a la que ya hemos hecho referencia, la aplicación rutinaria de los resultados algorítmicos generados automáticamente sin influencia real por el agente humano que las supervisa, no puede considerarse intervención significativa[596]. Para Veale y Edwards, ello implica considerar la frecuencia con la que el operador está en desacuerdo con el sistema y modifica o mejora los resultados[597].

En segundo lugar, para que la participación humana sea significativa, ésta debe llevarse a cabo por parte de una persona autorizada y competente para modificar la decisión[598], cuyo análisis debe tener en cuenta todos los datos pertinentes[599]. Brkan considera que tener en cuenta todos los datos pertinentes presenta una gran complejidad en la práctica, exacerbada en contextos en que las decisiones se toman a partir de correlaciones automáticas extraídas en contextos de *Big Data*[600].

Por otro lado, las directrices resaltan una diferencia para la intervención humana como medida de salvaguarda para decisiones basadas únicamente en el tratamiento automatizado; en este caso, el agente humano es calificado como revisor y su análisis debe tener en cuenta cualquier información adicional facilitada por el interesado[601]. Ello pone de manifiesto una vinculación instrumental entre la intervención humana como

[596] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 23. Parece implícita una alusión al sesgo de automatización, sobre la que profundizaremos al hablar del fundamento de la intervención humana en el RGPD. Vid. Apartado 1.2.2. Fundamento para la intervención humana en el artículo 22 RGPD, en este mismo capítulo.

[597] Veale y Edwards, «Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling», 400.

[598] En términos análogos, el considerando 48 de la propuesta de Reglamento AIA recoge que las personas físicas a quienes se haya encomendado la vigilancia humana deben poseer las competencias, la formación y la autoridad necesarias para desempeñar esa función. Sin embargo, este considerando no se traduce en una obligación para los usuarios de sistemas de IA en el artículo 29 AIA.

[599] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 23 y 30.

[600] En sus propias palabras: *In particular, it remains unclear how a human with limited capacities of data analysis will be able to justify that the final decision needs to be different from an algorithmic one, given that the automated system might not only have taken into account the data relating to the data subject affected by the decision, but a multitude of other complex datasets. If the automated decision was a simple sum of data appertaining to a particular data subject, an in-depth human review of automated decision would be much more feasible. If, however, the decision is based on complex relations between data in a Big Data environment, the human will have a much more difficult task in reviewing such a decision*. Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 108.

[601] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 30.

medida de salvaguarda con los derechos del interesado a una explicación, a expresar su punto de vista e impugnar la decisión[602].

En resumen, las directrices refrendadas por el CEPD definen que la intervención humana significativa debe [a] llevarse a cabo por persona autorizada y competente para modificar la decisión, [b] realizarse sobre un análisis que tenga en cuenta todos los datos disponibles y [c] no conllevar aplicación rutinaria de los resultados algorítmicos. Cuando esta participación humana se introduzca para el uso sistemas de apoyo a la toma de decisiones [2], dicha intervención significativa debe tener lugar necesariamente antes de la producción de efectos jurídicos o significativos para la persona interesada. Al contrario, cuando la participación humana se introduzca como medida de salvaguarda para las decisiones totalmente automatizadas [3], la intervención significativa tendrá lugar a requerimiento del interesado tras la producción de efectos jurídicos o significativos para el mismo, y la revisión debe, además, [d] tener en cuenta la información facilitada por el interesado.

Sin desmerecer esta labor interpretativa realizada por el GT29, cuyos elementos pueden considerarse una digna primera tentativa teórica para abordar el derecho a la intervención humana en el RGPD, no son, desde luego, suficientes para paliar la incertidumbre derivada de la escasa aplicación jurisprudencial, así como de la poca atención que la doctrina ha prestado a este elemento en particular[603].

1.2.2. Fundamento para la intervención humana en el artículo 22 RGPD

Con el objetivo de profundizar en esos elementos o criterios que conforman la intervención humana significativa en el artículo 22 del Reglamento, en este apartado se indaga en los fundamentos y razones para la incorporación de dicha disposición y su precedente en la normativa europea de protección de datos, aportando argumentos de carácter teleológico, en definitiva.

Lamentablemente, los trabajos preparatorios del RGPD, más centrados en la elaboración de perfiles y sus efectos discriminatorios, arrojan poca luz sobre la justificación del

[602] Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems», 1; Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 87; Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations», 22.

[603] Huq, «A Right to a Human Decision», 624.

artículo 22 RGPD[604]. No obstante, sí encontramos algunas pistas en su precedente inmediato, artículo 15 DPD, en el que se expresaron los miedos sobre el futuro de la dignidad humana ante el determinismo tecnológico[605].

Siguiendo el trabajo realizado por Mendoza y Bygrave, si atendemos al procedimiento legislativo para la inclusión del artículo 15 DPD, podemos observar que la preocupación general por el tratamiento automatizado como única lógica tras la toma de decisiones puede explicarse desde dos perspectivas complementarias. Por un lado, hay un temor a la pérdida del control humano en las decisiones que uno mismo toma, como responsable del tratamiento; por otro lado, hay un temor a esa pérdida de control humano en las decisiones que se toman sobre uno mismo, es decir, la pérdida de control sobre las decisiones que afectan a una de forma significativa, como persona interesada. Veamos detalladamente cómo se hacen explícitas estas dos perspectivas en el temor a la pérdida de control sobre la toma de decisiones, y cómo los distintos mecanismos de intervención humana cumplen con su propio rol normativo en este sentido.

En la primera versión de 1990, la Comisión pone de manifiesto ese temor Kafkiano relacionado con la pérdida de control sobre las decisiones que afectan a uno como interesado: «*Con esta disposición se pretende proteger el interés del interesado en participar en aquellas decisiones que sean importantes para él. El uso de perfiles detallados basados en datos personales por parte de importantes instituciones públicas y privadas priva al interesado de la posibilidad de influir en los procesos decisorios de dichas instituciones cuando esas decisiones se toman únicamente sobre la base de su perfil personal*»[606]. Es decir, la Comisión quiere con esta disposición que las personas interesadas participen en las decisiones que les afectan significativamente cuando se toman a partir del tratamiento de sus datos personales. Según Mendoza y Bygrave, el creciente uso de las técnicas de elaboración de perfiles, que provocaba que las personas

[604] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 83.

[605] Bygrave, «Minding the machine: art 15 of the EC Data Protection Directive and automated profiling», 18. Es relevante mencionar esa diferencia entre las culturas normativas estadounidense y europea, en lo que se refiere a la intervención humana ante los sistemas automatizados, dado que en esta última el respeto por la dignidad humana como valor normativo adquiere una relevancia esencial que es prácticamente desconocida en la cultura estadounidense. Vid. Huq, «A Right to a Human Decision», 629.

[606] Comisión de las Comunidades Europeas, Comunicación de la Comisión «sobre la protección de las personas en lo referente al tratamiento de datos personales en la Comunidad y a la seguridad de los sistemas de Información». Bruselas 24.09.1990. COM(90) 314 final - SYN 288, p. 22. Disponible aquí: https://eur-lex.europa.eu/procedure/EN/1990_287

tuvieran cada vez menos control y capacidad de influencia sobre las decisiones que les afectaban, fue el primer catalizador para la inclusión de ese artículo 15 DPD[607].

Tal y como se destaca en el apartado anterior, el derecho a obtener intervención humana bajo requerimiento del interesado está vinculado con los derechos del interesado a una explicación, a expresar su punto de vista e impugnar la decisión. Esta vinculación se fundamenta en el aspecto instrumental de la transparencia sobre el que disponer de posibilidades de ejercer los demás derechos reconocidos en el RGPD[608]. Este argumento lo recoge Malgieri al destacar cómo el Conseil Constitutionnel francés consideró la intervención humana como una salvaguarda fundamental en el diseño y desarrollo de algoritmos de IA[609], reconociendo ese vínculo entre el control humano y la capacidad de explicar, de forma detallada e inteligible, cómo se ha llevado a cabo el tratamiento a las personas interesadas[610].

Siguiendo esta misma línea, la doctrina ha destacado cómo los agentes humanos pueden tener la capacidad de ofrecer un rol intermediario entre la información que proporciona el sistema y la recibida por el interesado. Hamon et al. afirman que en los casos en los que no es fácil llegar a explicaciones satisfactorias podría ser útil una explicación mediada por el ser humano[611], si bien, Koivisto señala que la intermediación humana puede manejar la legitimidad en la percepción de la decisión sin corresponder dicha apariencia de legitimidad, necesariamente, con la realidad[612].

---

[607] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 83.

[608] Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 47.

[609] Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations», 15.

[610] Sentencia 2018-765 de 12 de junio de 2018 del Consejo Constitucional de Francia [Conseil Constitutionnel, Décision n° 2018-765 DC du 12 juin 2018], par. 71.

[611] Hamon et al., «Impossible Explanations? Beyond Explainable AI in the GDPR from a COVID-19 Use Case Scenario», 558.

[612] En sus propias palabras: *The more human mediation there is, resulting in carefully managed visibilities, the more legitimacy may be produced. At the same time, this may also mean less "truth", when the intricacies of the black box cannot, by being exposed, necessarily communicate anything (the truth-legitimacy trade-off).* Añade, y no debe olvidarse que, al margen del derecho de intervención humana, en el ejercicio de los derechos de información y acceso siempre concurre una intermediación humana que es relevante para el análisis jurídico. Koivisto, «Thinking Inside the Box: The Promise and Boundaries of Transparency in Automated Decision-Making», 19.

Volviendo al RGPD, podemos afirmar que la intervención humana en el apartado 22(3) tiene por objeto solicitar una segunda resolución, una revisión, en la que un agente humano puede tener en cuenta también el punto de vista del interesado[613]. Ello se reafirma en las directrices del GT29, al observar esa distinción entre los mecanismos de intervención humana, dado que el revisor humano debe tener en cuenta cualquier información adicional facilitada por el interesado cuando éste ejerce su derecho a la intervención humana para las decisiones basadas únicamente en el tratamiento automatizado [3][614]. En este caso, lo significativo de la intervención humana estaría directamente vinculado con la posibilidad del interesado de ejercer sus derechos a expresar su punto de vista y a impugnar la decisión[615].

Ahora bien, ya en esta primera versión de la Comisión en 1990 aparece un elemento que tendrá continuidad en el artículo 22 RGPD y se ha convertido en uno de los aspectos más endebles de esta regulación: «*El uso de perfiles (…) priva al interesado de la posibilidad de influir (…) cuando esas decisiones se toman únicamente sobre la base de su perfil personal*». La posibilidad de influir por parte de la persona interesada en estos procesos a través de medidas de salvaguarda como las descritas en el considerando 71[616], se incluyen exclusivamente para decisiones y perfiles basados únicamente en el tratamiento automatizado, excluyendo dicha influencia por parte del interesado de los sistemas de apoyo a la toma de decisiones.

A pesar de la importancia de esta perspectiva de pérdida de control por parte de los interesados sobre las decisiones que les afectan, en 1992 la Comisión Europea hizo referencia a una segunda perspectiva del temor a perder el control sobre la toma de decisiones automatizada: «*La utilización abusiva de la informática en la toma de decisiones constituye uno de los riesgos esenciales que se plantean en el futuro, ya que*

[613] Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems», 1; Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations», 22.

[614] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 30. Mientras que la información facilitada por el interesado no es relevante a la hora de definir la intervención humana significativa para los sistemas de apoyo a la toma de decisiones [2]

[615] Sobre este temor a la pérdida de control por parte de los interesados sobre las decisiones que les afectan significativamente y la conexión entre la intervención humana y el resto de derechos y medidas de salvaguarda volveremos más adelante.

[616] El derecho a obtener intervención humana – *out of the loop* –, a expresar su punto de vista, a recibir una explicación de la decisión tomada después de tal evaluación y a impugnar la decisión.

el resultado que proporciona la máquina, que utiliza programas cada día más refinados, e incluso sistemas expertos, tiene un carácter aparentemente objetivo e incontestable, al que el encargado de tomar las decisiones puede conceder una importancia excesiva, en dejación de su propia responsabilidad. (…) sólo se prohíbe la aplicación por parte del usuario de los resultados producidos por el sistema. La informática puede servir de ayuda a la decisión, pero no constituir en ningún caso su única base, debiendo dejarse a la apreciación humana el lugar que le corresponde»[617]. Efectivamente, en un mundo kafkiano en el que las grandes corporaciones adoptan constantemente decisiones automatizadas, no solo se desprovee de remedios efectivos a quienes se ven afectados por dichas decisiones, sino que también hay una pérdida de control y capacidad para explicar sus propias decisiones por parte de quienes las toman[618].

Hay varios elementos interesantes en estas líneas expuestas por la Comisión.

En primer lugar, la relación entre esa pérdida de control y la dejación de la responsabilidad hace referencia inevitablemente al principio de responsabilidad contenido en la normativa europea de protección de datos. Conforme al artículo 5 RGPD, en el tratamiento de datos personales el responsable del tratamiento debe cumplir y demostrar el cumplimiento con el Reglamento y sus principios. Así, el Reglamento introduce la intervención humana como un componente esencial de la toma de decisiones automatizada, con un mecanismo de gobernanza basado en la supervisión humana que impide la dejación de responsabilidad del responsable del tratamiento sobre las decisiones que adopta y afectan significativamente a las personas interesadas.

De hecho, la doctrina ha vinculado la intervención humana con el cumplimiento de varios de los principios del Reglamento, especialmente en lo que se refiere a la licitud y lealtad (art. 5(1)(a)) y a la exactitud (art. 5(1)(d)) en la toma de decisiones. Los humanos se consideran cruciales para evitar correlaciones indebidas y, por tanto, para garantizar la equidad en el tratamiento de datos[619], y no solo para eliminar la discriminación, sino

[617] Comisión de las Comunidades Europeas, Propuesta modificada de Directiva « relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos». Bruselas 15.10.1992. COM(92) 422 final- SYN 287, p. 27. Disponible aquí: https://eur-lex.europa.eu/procedure/EN/1990_287

[618] Korff, «Comments on Selected Topics in the Draft EU Data Protection Regulation», 25.

[619] Favaretto, De Clercq, y Elger, «Big Data and discrimination: perils, promises and solutions. A systematic review», 21.

también para reducir los falsos positivos[620]. Por su parte, Brkan sostiene que el artículo 22 del GDPR refleja el escepticismo europeo hacia los sesgos algorítmicos y las decisiones eventualmente erróneas que pueden tomar las máquinas no verificadas por los humanos[621]. En este sentido, para Hoofnagle, van der Sloot y Zuiderveen Borgesius, el considerando 71 es un mandato explícito para minimizar el riesgo de errores, tanto en términos de exactitud como de los posibles efectos discriminatorios[622]. Y las directrices del GT29 resaltan que los responsables del tratamiento deben tener en cuenta la exactitud en todas las fases del proceso de elaboración de perfiles, inclusive al aplicarlo para tomar una decisión que afecta a una persona[623].

De esta forma, la intervención humana como componente esencial de la toma de decisiones será significativa en la medida en que contribuya a un tratamiento automatizado lícito, leal y exacto, cuestión que el responsable del tratamiento deberá evaluar de forma apropiada[624].

Ahora bien, hay un segundo aspecto que debe destacarse en esta perspectiva, la Comisión expresa su preocupación por conceder una importancia excesiva a los resultados automatizados, y añade que la informática puede servir de ayuda, pero nunca constituir la única base de la toma de decisiones. Podemos identificar aquí un temor al determinismo tecnológico que pone en valor, a su vez, el juicio humano.

---

620 Roig, «Safeguards for the right not to be subject to a decision based solely on automated processing (Article 22 GDPR)», 6.

621 Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 97. También vemos esta idea en relación con la regulación de la IA y la inclusión de la supervisión humana como requisito obligatorio. En el Libro Blanco sobre IA se entiende que entiende que la supervisión humana contribuye a garantizar que un sistema de IA no cause efectos adversos, vid. Comisión Europea, Libro Blanco sobre la inteligencia artificial, 21. Mientras que en la propuesta de Reglamento AIA, el artículo 14(2) establece que la supervisión humana tendrá como objetivo prevenir o minimizar los riesgos para la salud, la seguridad o los derechos fundamentales.

622 Hoofnagle, van der Sloot, y Borgesius, «The European Union general data protection regulation: what it is and what it means», 92.

623 Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 13.

624 Sobre si realmente la intervención humana puede contribuir a una toma de decisiones más lícita y exacta se discutirá más adelante, recogiendo también al sector doctrinal más crítico en este aspecto. Vid. Apartado 1. ¿Es posible una intervención humana significativa? Partiendo de las críticas a la intervención humana para reivindicar una supervisión humana basada en la evidencia, en Capítulo 4. La intervención humana y el principio de responsabilidad en el tratamiento de datos personales: un enfoque basado en la evidencia a través de la evaluación de impacto. Una propuesta desde la medicina preventiva.

Mendoza y Bygrave resaltan que la Comisión recoge en esta disposición no sólo el temor a que los humanos dejen que las máquinas cometan errores, sino la preocupación por mantener la dignidad humana garantizando que los humanos mantengan el papel principal en la "constitución" de sí mismos[625].

Ahora bien, han de diferenciarse aquí dos dimensiones distintas en esta concepción.

Por un lado, una versión agravada de este fundamento nos llevaría a la prohibición del tratamiento automatizado en sí, tanto si sirve como apoyo a la toma de decisiones o es su fundamento único, esta prohibición configuraría un espacio inviolable y no-computable de la personalidad frente a la automatización. Esta dimensión no parece tener cabida en el artículo 22 dado que, precisamente, permite cualquier clase de tratamiento siempre que no esté basada únicamente en el tratamiento automatizado[626].

Ello nos lleva a la segunda dimensión de esta concepción, una suerte de reserva de humanidad en la toma de decisiones[627]. En este mismo sentido, para Jones la dignidad

[625] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 84. Según Hildebrandt, deberíamos rechazar que los algoritmos pueden definir nuestro ser, puesto que éste no es computable *–the incomputable self–*: *We should resist attempts to lure us into accepting the drawbacks of "computer says no" based on a flawed belief in computers that supposedly "outperform" human decision-makers*. Hildebrandt, «Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning», 119.

[626] Así lo expresaba el Supervisor Europeo de Protección de Datos (SEPD) sobre la elaboración de perfiles en los trabajos preparatorios del RGPD: *El problema no es (…) la práctica de la elaboración de perfiles, sino, más bien, la falta de información adecuada sobre la lógica algorítmica a partir de la que se desarrollan tales perfiles y que repercute en el interesado*. Supervisor Europeo de Protección de Datos (SEPD), «Dictamen 3/2015 -La gran oportunidad de Europa. Recomendaciones del SEPD sobre las opciones de la UE en cuanto a la reforma de la protección de datos-». Sí parece encontrarse este fundamento, por contra, en el artículo L10 del *Code de justice administrative* francés, en su modificación por la *LOI n°2019-222 du 23 mars 2019*, en la que se incluye la siguiente prohibición: *Les données d'identité des magistrats et des membres du greffe ne peuvent faire l'objet d'une réutilisation ayant pour objet ou pour effet d'évaluer, d'analyser, de comparer ou de prédire leurs pratiques professionnelles réelles ou supposées. La violation de cette interdiction est punie des peines prévues aux articles 226-18,226-24 et 226-31 du code pénal, sans préjudice des mesures et sanctions prévues par la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*. A pesar de las múltiples ambigüedades que contiene, este fundamento también parece estar presente en las prohibiciones de uso de IA del artículo 5 de la propuesta de Reglamento AIA, por ejemplo, artículo 5(1): a) *La introducción en el mercado, la puesta en servicio o la utilización de un sistema de IA que se sirva de técnicas subliminales que trasciendan la conciencia de una persona para alterar de manera sustancial su comportamiento de un modo que provoque o sea probable que provoque perjuicios físicos o psicológicos a esa persona o a otra*.

[627] Ponce Solé, «Inteligencia artificial, Derecho administrativo y reserva de humanidad: algoritmos y procedimiento administrativo debido tecnológico», 28-30; Cotino, «Ética en el diseño y confiable para el desarrollo de la robótica, inteligencia artificial y el big data y su utilidad desde el derecho», 36. Sobre este fundamento es interesante desarrollado por parte de la doctrina que habla de la empatía entre los participantes en la toma de decisiones o *role-reversity*, esto es, la idea de que quienes ejercen el juicio deben ser recíprocamente vulnerables a sus procesos y efectos, como principio democrático fundamentado en la dignidad humana. Brennan-Marquez y Henderson, «Artificial Intelligence and Role-Reversible Judgment», 140; Morente Parra, «Big Data o el arte de analizar datos masivos. Una reflexión crítica desde los derechos

humana en la cultura normativa europea puede ser restaurada a través de la intervención humana, de forma que los humanos no sean tratados de forma exclusivamente computacional[628]. La intervención humana en la fase decisoria se erige aquí como remedio regulatorio al determinismo tecnológico. La inclusión de la intervención humana como componente esencial de la toma de decisiones en el apartado 22(1) respondería a esa preocupación por el determinismo tecnológico.

¿Cómo se integra esta dimensión en el RGPD?

Por un lado, dadas las amplias excepciones establecidas por el RGPD en los apartados 2 y 4, no puede decirse que se conciba en sí una vulneración de la autonomía y dignidad humana por la toma de decisiones basada únicamente en el tratamiento automatizado. Si encontrásemos alguna prohibición insalvable en el artículo 22 RGPD, imaginemos que el apartado 4 no admitiese excepciones para la toma de decisiones basada en categorías especiales de datos, podríamos concluir que el Reglamento concibe la ausencia de intervención humana como incompatible con la autonomía y dignidad humana en dicho espacio. Sin embargo, este no es el caso.

Por otro lado, ¿qué reflejo tiene esta dimensión en la significancia de la intervención humana como componente esencial de la toma de decisiones? Al analizar las directrices del GT29 hemos podido resaltar cómo la aplicación rutinaria de los resultados algorítmicos generados automáticamente sin influencia real por el agente humano que las supervisa no puede considerarse intervención significativa[629]. Por ende, intervención

fundamentales», 240. No deja de ser un argumento de interés en el debate jurídico a pesar de que no parece tener cabida en el RGPD. Dos puntualizaciones a este respecto. La primera sería que, como bien indican Brennan-Marquez y Henderson, la caracterización "humana" de este principio no impide que una mayoría de los sistemas institucionales de toma de decisiones humanos existentes no satisfagan el criterio normativo en la práctica. Vid. Brennan-Marquez y Henderson, «Artificial Intelligence and Role-Reversible Judgment», 143. En cualquier caso, es positivo que los debates sobre las tecnologías de IA nos obliguen a replantear cuestiones que podemos dar por sentadas en el ejercicio del juicio humano. Por otro lado, el hecho de que no participen agentes humanos capaces de empatizar en la fase de toma de decisiones, no quiere decir que no lo hagan en fases de diseño y desarrollo o en el despliegue de estas tecnologías. Estos agentes humanos deberían ser igualmente conscientes y empáticos con las repercusiones que sus decisiones tienen sobre otros humanos en el uso de las tecnologías.

[628] Jones, «The right to a human in the loop: Political constructions of computer automation and personhood», 232.

[629] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 23. También encontramos esta idea en lo que configura una supervisión humana efectiva para la Comisión en la propuesta de Reglamento AIA, entre los requisitos del artículo 14, los sistemas de IA deben ser desarrollados de tal forma que la personas a las que se asigne la supervisión puedan: *ser conscientes de la posible tendencia a*

humana significativa es aquella que no concede una importancia *excesiva* a los resultados automatizados. Es decir, el responsable del tratamiento debe evaluar si las personas que, bajo su autoridad, han sido encomendadas para intervenir en el proceso decisorio, incurren o no en ese sesgo de automatización o, en definitiva, aplican de forma rutinaria dichos resultados sin influencia real en la toma de decisiones.

En definitiva, el vistazo a los trabajos preparatorios del artículo 15 DPD permite vislumbrar algunas de las razones por las cuales la intervención humana fue incluida en el mismo y en el actual artículo 22 RPGD.

Por un lado, en lo que respecta a la intervención humana como medida de salvaguarda para la toma de decisiones basada únicamente en el tratamiento automatizado -22(3)-, la Comisión pone de manifiesto la preocupación por la pérdida de control por parte de los interesados en las decisiones que les afectan de forma directa. Ello ahonda en ese vínculo entre la intervención humana y el resto de las medidas de salvaguarda y, por ende, una intervención significativa que responda a esta preocupación será aquélla que facilite al interesado influir sobre la decisión adoptada.

Por otro lado, en lo que se refiere a la intervención humana como componente esencial de la toma de decisiones -22(1)-, la Comisión expresa un temor por la pérdida de control sobre las decisiones que los mismos responsables toman y afectan de forma directa a las personas interesadas. De esta forma, la intervención humana se integra en el RGPD para responsabilizar al responsable del tratamiento de su cumplimiento[630]. Por tanto, si la intervención humana contribuye, en particular, al cumplimiento de los principios de licitud, lealtad y exactitud en el tratamiento será significativa, salvo que dicha intervención se limite a aplicar rutinariamente los resultados algorítmicos concediendo una importancia excesiva a los mismos.

2. Derecho a la información en la toma de decisiones automatizada

confiar automáticamente o en exceso en la información de salida generada por un sistema de IA de alto riesgo («sesgo de automatización») -art. 14(4)(b) AIA-.

[630] Este fundamento en el principio de responsabilidad hace tan relevante la conexión de la intervención humana como elemento esencial en la toma de decisiones con la herramienta fundamental en el RGPD para el cumplimiento de este principio, es decir, la evaluación de impacto de protección de datos. Vid. Capítulo 4. La intervención humana y el principio de responsabilidad en el tratamiento de datos personales: un enfoque basado en la evidencia a través de la evaluación de impacto. Una propuesta desde la medicina preventiva.

Durante la tramitación legislativa del RGPD, en su Dictamen 3/2015, el Supervisor Europeo de Protección de Datos (SEPD) expresó su visión y recomendaciones respecto de la normativa de protección de datos. Entre otras, al referirse a la elaboración de perfiles y la toma de decisiones a partir de éstos, recomendaba un mayor grado de transparencia por parte de los responsables del tratamiento en los siguientes términos: «*El problema no es (...) la práctica de la elaboración de perfiles, sino, más bien, la falta de información adecuada sobre la lógica algorítmica a partir de la que se desarrollan tales perfiles y que repercute en el interesado*»[631]. Esta expresión no es más que uno de los muchos ejemplos de esta visión que es compartida en la doctrina que analiza la protección de datos en el ámbito europeo. Consecuencia de ello, desde la aprobación del Reglamento, los derechos de información en aplicación del principio de transparencia han sido protagonistas ineludibles del debate doctrinal sobre el artículo 22 RGPD.

Por relevante que sea esta visión, el principio de transparencia contiene limitaciones severas que se expondrán a continuación y, como consecuencia, encorseta el alcance de dicha disposición y del Reglamento y el resto de sus principios. En esta investigación, trato de contrarrestar esta visión, aunque no por ello se deja de poner en valor la misma como un pilar esencial en la interpretación del artículo 22 RGPD. Antes de ello, las limitaciones de la transparencia como principio normativo.

La falacia de la transparencia es un término habitualmente utilizado en la doctrina para ilustrar cómo ésta puede, no solo demostrarse ineficaz para los fundamentos normativos que la motivan, sino incluso resultar perjudicial para los mismos.

Hoepman presenta varias razones para no caer en esta falacia. En primer lugar, la transparencia es inútil cuando el desequilibrio de poder entre el responsable del tratamiento y la persona interesada es tal, que la capacidad operativa o agencia de esta última es inexistente[632]. Además, los estándares normativos de la transparencia, en el caso que nos ocupa los descritos en el artículo 12 RGPD, pueden ser difíciles de cumplir cuando se utilizan algoritmos complejos cuyos procesos internos son ininteligibles para

[631] Supervisor Europeo de Protección de Datos (SEPD), «Dictamen 3/2015 -La gran oportunidad de Europa. Recomendaciones del SEPD sobre las opciones de la UE en cuanto a la reforma de la protección de datos-», 10.

[632] Hoepman, «Transparency is the perfect cover-up (if the sun does not shine)», 47.

los humanos[633]. Y no menos importante, impugnar una decisión puede ser imposible sobre la base de una explicación, que incluso siendo válida y razonable en términos normativos podría no ser el verdadero fundamento de la misma[634].

En definitiva, el principio de transparencia es un requisito necesario, pero no suficiente para el desarrollo y uso de sistemas algorítmicos que garanticen el respeto de los derechos y libertades del interesado, tampoco para cerrar esa brecha entre desarrolladores, clientes, reguladores y ciudadanía en general[635].

Estas limitaciones también han sido señaladas en lo que se refiere, particularmente, a los derechos de información aplicables a la toma de decisiones automatizada en el RGPD, en el contexto de la discusión doctrinal sobre el alcance del derecho a una explicación.

Cobbe y Singh destacan que, como tal, las explicaciones sobre cómo un modelo llega a una decisión particular obvian los aspectos más relevantes de la toma de decisiones y, desde un punto de vista normativo, dichas explicaciones no proveen de la información necesaria para determinar si se ha llegado a una decisión de forma legítima, o si resultan discriminatorias[636]. Para Hamon et al. tanto el apartado tercero del artículo 22 como el considerando 71, al abordar las medidas de salvaguarda para los derechos de los interesados no se centran en el derecho a una explicación, sino en una variedad de herramientas contenidas en el RGPD, destacando entre las obligaciones del responsable del tratamiento para cumplir y demostrar el cumplimiento en la toma de decisiones automatizada basada en la elaboración de perfiles, como los principios de licitud y lealtad, el enfoque basado en el riesgo o el modelo de Evaluación de la EIPD[637].

---

633 Hoepman, 47.

634 Hoepman, 48. Tal y como señala Koivisto, la intermediación humana puede manejar la legitimidad en la percepción de la decisión sin corresponder dicha legitimidad, necesariamente, con la realidad. Vid. Koivisto, «Thinking Inside the Box: The Promise and Boundaries of Transparency in Automated Decision-Making», 19. En otras ocasiones, el análisis completo de la decisión podría estar exclusivamente en manos del responsable del tratamiento incluso por razones legítimas, por ejemplo, por limitar el análisis de datos personales que son de terceros e influyen de forma determinante en la decisión que afecta a otra persona interesada.

635 Berscheid y Roewer-Despres, «Beyond Transparency: A Proposed Framework for Accountability in Decision-Making AI Systems».

636 Cobbe y Singh, «Reviewable Automated Decision-Making», 2.

637 Hamon et al., «Impossible Explanations? Beyond Explainable AI in the GDPR from a COVID-19 Use Case Scenario», 558.

No obstante, a pesar de sus limitaciones, la transparencia no deja de ser un principio fundamental dentro del ecosistema regulatorio del RGPD y un pilar esencial en la regulación de las decisiones automatizadas. Antes, entre los fundamentos de la disposición kafkiana, se destacaba el control por parte de las personas interesadas sobre las decisiones que les afectan de forma significativa. A continuación, veremos que el cumplimiento del principio de transparencia es de vital importancia para garantizar el derecho a impugnar las decisiones automatizadas y, en definitiva, para garantizar el (limitado) control que los interesados tienen sobre la adopción de decisiones sobre la base del tratamiento de sus datos personales[638].

No obstante, los derechos de información y acceso que el RGPD garantiza para la toma de decisiones automatizada es diferente en función de si ésta está basada o no únicamente en el tratamiento automatizado. Veámoslo con detalle.

2.1. Derechos de información sobre las inferencias algorítmicas en el RGPD

Los artículos 13(2)(f), 14(2)(g) y 15(1)(h) RGPD establecen una serie de derechos de información y acceso específicos para la toma de decisiones automatizada, sin embargo, el objetivo de este apartado es determinar de forma sucinta cuáles son los derechos de información aplicables al estadio anterior, esto es, a las inferencias algorítmicas o elaboración de perfiles sobre las que se basa posteriormente la toma de decisiones automatizada[639].

Si distinguimos entre datos de entrada -*inputs*- o datos en bruto de carácter personal y datos de salida -*outputs*- que serían los datos inferenciales que el tratamiento arroja, podemos adelantar que las garantías jurídicas que el RGPD ofrece para esta última clase de datos personales resultan insuficientes: «*[i]ronically, inferences receive the least protection of all the types of data addressed in data protection law, and yet now pose perhaps the greatest risks in terms of privacy and discrimination*»[640]. Son dos cuestiones básicas las que debemos aclarar para abordar esta cuestión, primero, si dichas inferencias

[638] Garriga Domínguez, «La elaboración de perfiles y su impacto en los derechos fundamentales. Una primera aproximación a su regulación en el reglamento general de protección de datos de la Unión Europea», 135.

[639] Volviendo sobre lo dicho anteriormente, no toda elaboración de perfiles -art. 4(4) RGPD- ha de ser considerada como una decisión automatizada individual -art. 22 RGPD-.

[640] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI», 575.

pueden ser consideradas datos personales, segundo, si pueden considerarse datos personales, ha de determinarse el alcance de los derechos de información y acceso, así como del principio de transparencia sobre los mismos.

Parece pacífico considerar que los datos inferenciales y, por ende, los perfiles deben considerarse datos personales. Para el CEPD, a partir de dicha inferencia o perfilado el responsable del tratamiento genera datos personales "nuevos" que no han sido directamente facilitados por los propios interesados[641]. A la hora de considerar los elementos que configuran los datos personales[642], el elemento clave para las inferencias o perfiles es si puede considerarse información "sobre" una persona[643]. Korff expone que, aunque a primera vista alguien pudiera pensar que un perfil es simplemente un indicador genérico resultante de una serie de suposiciones estadísticas, quizá derivados de datos sobre personas, pero no relacionados como tales con ninguna persona identificable, no pueden obviarse los efectos de la aplicación de dichos perfiles (Korff 2010, p. 52)[644]. Es decir, con independencia de que los datos utilizados en el "perfil" se refieran directamente, en su contenido, a las personas, están (i) específicamente destinados a ser aplicados a las personas, y (ii) pueden tener efectos extremadamente graves, incluso devastadores, sobre las personas que resulten "coincidentes" -total o parcialmente, añadiría-[645]. Además, independientemente de los datos de entrada que se hayan utilizado,

[641] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 10. Lo cual implica que dichos datos deben informarse conforme al artículo 14(3) RGPD: *3. El responsable del tratamiento facilitará la información indicada en los apartados 1 y 2: a) dentro de un plazo razonable, una vez obtenidos los datos personales, y a más tardar dentro de un mes, habida cuenta de las circunstancias específicas en las que se traten dichos datos; b) si los datos personales han de utilizarse para comunicación con el interesado, a más tardar en el momento de la primera comunicación a dicho interesado, o c) si está previsto comunicarlos a otro destinatario, a más tardar en el momento en que los datos personales sean comunicados por primera vez.*

[642] Vid. al respecto Romeo Casabona, «Datos personales (Comentario al artículo 4. 1 RGPD)».

[643] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI», 518.

[644] Korff, «New Challenges to Data Protection Study - Comparative Chart: Divergencies between Data Protection Laws in the EU», 52. Korff sigue aquí el modelo de 3 elementos propuesto por el GT29: para considerar que los datos versan «sobre» una persona debe haber un elemento «contenido» o un elemento «finalidad» o un elemento «resultado»; interpretación también adoptada por el TJUE en el caso Nowak: *(…) Este último requisito se cumple cuando, debido a su contenido, finalidad o efectos, la información está relacionada con una persona concreta*. STJUE de 20 de diciembre de 2017, caso *Peter Nowak contra Data Protection Commissioner* (Asunto C-434/16), par. 35.

[645] Korff, 53.

dichos datos inferenciales pueden ser considerados de categoría especial cuando revelen información recogida en el artículo 9(1) RGPD[646].

Otra cuestión discutida es si el hecho de que dichos datos hayan sido "generados" de forma algorítmica por el responsable del tratamiento puede suponer o no un obstáculo para considerar que sean datos del interesado. Y no puede serlo. Incluso las opiniones "generadas" por terceros -humanos- son consideradas datos personales de aquellas personas a las que se refieren, dada la interpretación del TJUE en el caso *Nowak*[647] sobre las anotaciones subjetivas en los exámenes[648]. El TJUE determina que dichas anotaciones han de ser consideradas como datos personales, lo cual no obsta para que la misma información pueda considerarse dato personal de otra persona, en este caso de la persona examinadora[649]. Haciendo una interpretación analógica de esta argumentación, si el hecho de que una información que es el resultado de un proceso cognitivo llevado a cabo por un tercero -humano- puede considerarse dato personal respecto de aquella persona a la que se refiere la información (cumpliendo con los elementos definidos anteriormente), no hay razón para considerar que una información resultado de una inferencia algorítmica no pueda ser igualmente considerada como dato personal respecto de aquella persona a la que se refiera.

En definitiva, puede considerarse que los datos inferidos o perfiles son datos personales, más problemático es discutir sobre el alcance de los derechos de información y acceso sobre dichos datos. A mi entender, debemos distinguir tres dimensiones en lo que se refiere a la elaboración de perfiles (1) la información sobre la propia actividad de

---

[646] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI», 515. En mi opinión, de forma analógica al razonamiento expuesto anteriormente, aunque la inferencia o perfil no revele necesariamente un contenido de categoría especial, ello no obsta para que pueda considerarse un dato personal de categoría especial cuando su aplicación, sea por su finalidad o sus efectos, tenga un impacto inequívoco sobre alguna de las categorías recogidas en el artículo 9(1) RGPD.

[647] STJUE de 20 de diciembre de 2017, caso *Peter Nowak contra Data Protection Commissioner* (Asunto C-434/16). En dicho caso el tribunal debía decidir sobre el alcance de los derechos de acceso – y de la aplicación del resto de derechos en la DPD – sobre el examen que había realizado el Sr. Nowak, incluyendo sus respuestas (que fueron consideradas datos personales), las preguntas (que no fueron consideradas tales) y las anotaciones de las personas examinadoras (cuestión más polémica y sobre la que entramos a continuación). Vid. al respecto Jove, «Peter Nowak v Data Protection Commissioner».

[648] El TJUE define las mismas como: *"la opinión o valoración del examinador sobre los resultados […] y, en particular, sobre sus conocimientos y competencias" con la finalidad de "documentar la evaluación de los resultados del aspirante"*. STJUE de 20 de diciembre de 2017, caso *Peter Nowak contra Data Protection Commissioner* (Asunto C-434/16), par. 43.

[649] Jove, «Peter Nowak v Data Protection Commissioner», 177.

perfilado entre los fines del tratamiento; (2) el perfil generado, esto es, el dato personal en sí y; (3) la lógica y consecuencias de dicha actividad.

Tanto si los datos personales de entrada -*inputs*- para la realización del perfil se obtienen del interesado, como si no, existe la obligación de informar sobre los fines del tratamiento a que se destinan los datos personales -arts. 13(1)(c) y 14(1)(c) RGPD-, esto es, se debe comunicar a la persona interesada en todo caso que sus datos personales van a ser objeto de un tratamiento con fines de elaboración de perfiles[650]. El derecho de acceso incluye también esa primera dimensión sobre los fines de perfilado del tratamiento de datos personales.

Sobre el propio perfil generado -*output*-, Wachter y Mittelstadt argumentan que hay una laguna en el RGPD que permite a los responsables del tratamiento eludir la obligación de informar extrayendo ellos mismos los datos inferidos. Esto ocurre, según estos autores, en los casos en que los datos inferidos o derivados no se obtienen a través de un tercero, sino que son creados por el propio responsable del tratamiento, porque el Reglamento solo obliga a notificar cuando se obtienen directamente del interesado, artículo 13 RGPD, o de un tercero, artículo 14 RGPD[651]. Desde mi punto de vista, el artículo 14 no establece que haya necesariamente un tercero que comunique esos datos, sino que simplemente los datos personales no se obtengan directamente de la persona interesada. En cualquier caso, lo relevante aquí es que el derecho de acceso sí incluye ese derecho a conocer el dato de salida del sistema -*output*-, o perfil[652], dado que debe considerarse un dato personal como tal, conforme a los argumentos arriba expuestos, pudiendo obtener del responsable el derecho de acceso a los datos personales que le conciernen -art. 15(1) RGPD-. Si, además, consideramos que en todo caso el responsable del tratamiento debe informar sobre la

[650] En este mismo sentido, las Directrices del GT29 sobre elaboración de perfiles y decisiones automatizadas argumentan que la información sobre los fines de perfilado debe facilitarse en todo caso, aunque dicho perfilado no entre en el ámbito del artículo 22 RGPD: *En particular, cuando el tratamiento implique la toma de decisiones basada en la elaboración de perfiles (independientemente de si entran en el ámbito de las disposiciones del artículo 22), debe aclararse al usuario el hecho de que el tratamiento tiene fines tanto de a) elaboración de perfiles como de b) adopción de una decisión sobre la base del perfil generado*. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 18.

[651] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI», 545.

[652] Así lo ven también Wachter y Mittelstadt con apoyo en varias de las Directrices del GT29, en Wachter y Mittelstadt, 545.

propia actividad de perfilado, primera dimensión, la persona interesada estará en todo caso legitimada para el acceso al perfil generado, segunda dimensión.

Por último, ha de valorarse el alcance de estos derechos sobre la lógica y consecuencias de dicha actividad de perfilado. Los Considerando 60 y 63 RGPD parecen sugerir, primero, que no hay un mero derecho de acceso sino un deber de informar sobre la existencia de perfiles[653], y segundo, que los derechos de información y acceso cuando se trata de elaboración de perfiles van más allá de la información sobre la existencia del perfil en sí -segunda dimensión-: «*(…) Se debe además informar al interesado de la existencia de la elaboración de perfiles y de las consecuencias de dicha elaboración*». También se reitera en el Considerando 63 RGPD: «*(…) Todo interesado debe, por tanto, tener el derecho a conocer y a que se le comuniquen, en particular, los fines para los que se tratan los datos personales, su plazo de tratamiento, sus destinatarios, la lógica implícita en todo tratamiento automático de datos personales y, por lo menos cuando se base en la elaboración de perfiles, las consecuencias de dicho tratamiento*».

No obstante, los artículos 13(2)(f), 14(2)(g) y 15(1)(h) RGPD al hablar de información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas de las decisiones automatizadas, incluida la elaboración de perfiles, para el interesado, hacen una referencia directa a la toma de decisiones basada únicamente en el tratamiento automatizado que produzca efectos jurídicos en él o le afecte significativamente de modo similar: «*la existencia de decisiones automatizadas, incluida la elaboración de perfiles, a que se refiere el artículo 22, apartados 1 y 4*». Ahora bien, es cierto que el legislador hace referencia a dicha clase de tratamiento, sí, pero también establece que deberá informarse al interesado o permitirle acceder a tal información *al menos en tales casos* -es decir, "al menos" en las decisiones basadas únicamente en el tratamiento automatizado, sin excluir otras posibilidades-.

Contamos cuando menos con un antecedente que sugiere que el derecho de acceso a información significativa sobre la lógica aplicada en la elaboración de perfiles, así como la importancia y las consecuencias previstas debe aplicarse más allá de la clase de tratamiento prohibido por los artículos 22(1) y (4) RGPD. La autoridad de control austríaca, *Datenschutzbehörde*, consideró que el derecho de acceso en el artículo 15(1)(h)

[653] Aunque el término "la existencia de la elaboración de perfiles" no deja claro si estamos hablando de esa primera o segunda dimensión arriba definidas.

RGPD se aplica a toda clase de elaboración de perfiles y no solo a la que produce efectos jurídicos o significativos basada únicamente en el tratamiento automatizado. En este caso[654], se discutía sobre el alcance de los derechos de información y acceso de una herramienta de marketing llamada *GeoMilieus* y que, según la información aportada por la propia compañía a la persona interesada, calificaba a las personas en grupos como "conservadores", "tradicionales", "pragmáticos" o "hedonistas" para ajustar sus preferencias de marketing.

La autoridad de control determina en primer lugar que dicho ajuste personalizado y probabilístico debía ser calificado como dato personal conforme al artículo 4(1) RGPD y elaboración de perfiles conforme al artículo 4(4) RGPD, por evaluar aspectos personales como la situación económica, o las preferencias o intereses personales[655]. Posteriormente, determina que la elaboración de perfiles realizada por *GeoMilieus*, aunque no sea subsumible en la elaboración de perfiles prohibida con carácter general por los artículos 22(1) y (4), debe garantizar el acceso a la información definida en el artículo 15(1)(h) puesto que el mismo hace referencia a que la información debe facilitarse *al menos en tales casos* –esto es, en el tratamiento prohibido generalmente por los artículo 22(1) y (4)–, lo cual obliga a una interpretación amplia de este derecho que debe extenderse a toda clase de elaboración de perfiles y toma de decisiones automatizada[656]. Por último, hace referencia también a las directrices del GT29 para sostener su argumento[657]. Estas directrices indican que, para la elaboración de perfiles, además de la información general sobre el tratamiento: «*el responsable del tratamiento tiene el deber de poner a disposición los datos utilizados como datos de entrada para crear perfiles, así como de facilitar el acceso a la información sobre el perfil y los detalles sobre los segmentos a los que se ha asignado al interesado*»[658].

---

[654] Resolución de la Autoridad de Control Austríaca (Datenschutzbehörde) núm. 2020-0.436.002, de 8 de septiembre de 2020 (ref.: DSB-D124.909). Un resumen y traducción al inglés de la resolución puede encontrarse en GDPRhub, aquí: https://gdprhub.eu/index.php?title=DSB_(Austria)_-_2020-0.436.002

[655] Resolución de la Autoridad de Control Austríaca (Datenschutzbehörde) núm. 2020-0.436.002, de 8 de septiembre de 2020, pár. D.1.b. y f.

[656] Resolución de la Autoridad de Control Austríaca (Datenschutzbehörde) núm. 2020-0.436.002, de 8 de septiembre de 2020, pár. D.1.h.

[657] Resolución de la Autoridad de Control Austríaca (Datenschutzbehörde) núm. 2020-0.436.002, de 8 de septiembre de 2020, pár. D.1.i.

[658] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 19.

La autoridad de control austríaca incide sobre una voluntad del legislador que, en realidad, no parece ofrecer dudas: *al menos en tales casos*, es una expresión clara de la voluntad del legislador de que dichos derechos de información y acceso no sean únicamente garantizados para la clase de tratamiento definida en el artículo 22(1) y (4) –basados únicamente en el tratamiento–[659], sin embargo, ¿cómo determinar el alcance del mismo ante una expresión tan abierta y ambigua?

Conforme al enfoque basado en el riesgo adoptado por el RGPD al que ya se ha hecho referencia, hay dos posibles interpretaciones que parecen compatibles a la hora de interpretar el alcance de la expresión "*al menos en tales casos*", sin expandir este derecho a toda clase de elaboración de perfiles como sugiere la autoridad de datos austríaca. Por un lado, estos derechos podrían aplicarse a toda clase de decisiones automatizadas –incluida la elaboración de perfiles– que produzcan efectos jurídicos o afecten de forma significativa a las personas interesadas conforme al artículo 22 RGPD. De esta forma, siempre que el responsable del tratamiento deba introducir intervención humana significativa antes de la producción de dichos efectos para evitar las prohibiciones del 22(1) y (4) RGPD, deberá igualmente garantizar los derechos 13(2)(f), 14(2)(g) y 15(1)(h) RGPD. Por otro lado, estos derechos podrían aplicarse a toda clase de decisiones automatizadas –incluida la elaboración de perfiles– que entre en el marco de la obligación de realizar una EIPD por parte del responsable del tratamiento, lo cual sería además coherente con el artículo 35 RGPD que no distingue, a la hora de determinar si determinado tratamiento es de riesgo alto para los interesados, entre decisiones o elaboración de perfiles basada únicamente o no en el tratamiento automatizado.

2.2. Derecho a la información para las decisiones basadas únicamente en el tratamiento automatizado, ¿derecho a una explicación?

En este apartado se abordará la cuestión más discutida en la literatura –no solo jurídica[660]– acerca de la toma de decisiones automatizada. Más arriba se han señalado las limitaciones

[659] No obstante, Malgieri y Comandé mantienen una visión mucho más restrictiva y sostienen que el alcance de esta cláusula es únicamente aplicable a tales casos, mientras que el RGPD abre a que voluntariamente los responsables del tratamiento lo adopten en otros casos, en Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 250.

[660] Como muestra pueden consultarse los programas de la ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) en los últimos años, desde su creación en 2018, los cuáles muestran como esta cuestión ha sido la más discutida de forma interdisciplinar. Disponible en: https://facctconference.org/index.html

de este enfoque, no tanto porque no sea útil por sí mismo, sino porque su protagonismo ha desplazado el análisis de otros mecanismos y principios de regulación de las decisiones automatizadas en el RGPD igualmente útiles. En definitiva, no es mi pretensión añadir nada nuevo a esta discusión, y por ello me limitaré a describir las posiciones más relevantes y a integrar dicha discusión en el enfoque de esta investigación.

Una de las causas que desencadenaron esta discusión la encontramos en el propio texto del RGPD. El Considerando 71 sobre la toma de decisiones, incluida la elaboración de perfiles, basada únicamente en el tratamiento automatizado y que produzca efectos jurídicos en él o le afecte significativamente de modo similar, dice así: «*En cualquier caso, dicho tratamiento debe estar sujeto a las garantías apropiadas, entre las que se deben incluir la información específica al interesado y el 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*». Mientras que el artículo 22(3) RGPD al regular dichas garantías dice: *el responsable del tratamiento adoptará las medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado, como mínimo el derecho a obtener intervención humana por parte del responsable, a expresar su punto de vista y a impugnar la decisión*». La doctrina pronto se preguntó qué había ocurrido con la salvaguarda del derecho a recibir una explicación de la decisión tomada.

Las Directrices del GT29 recogen que entre las garantías a incluir por el responsable del tratamiento está el derecho a una explicación (GT29 2018, 30), con lo que parecen sugerir que el Considerando 71 es jurídicamente vinculante, cuestión que fue criticada por la doctrina dado que ello contradice la jurisprudencia del TJUE[661].

A partir de aquí, se plantearon distintas posiciones, algunas defendiendo que el derecho a una explicación derivaba de los derechos a obtener "información significativa sobre la

[661] Los Considerando pueden ayudar a explicar el propósito y la intención de un instrumento normativo, o tenerse en cuenta para resolver ambigüedades en las disposiciones legislativas a las que se refieren, pero no son vinculantes en términos normativos de forma autónoma Vid. Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 254-55. Una interpretación más acorde a la jurisprudencia del TJUE propone Brkan, quien entiende que basarse en un enfoque de interpretación conjunta de varias disposiciones del RGPD, incluyendo los derechos de información y acceso y el artículo 22, y utilizar el considerando 71 para reforzar la interpretación que apoya la existencia del derecho del interesado a obtener las razones de la decisión automatizada, no conduciría a una interpretación *contra legem*, en Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 114.

lógica aplicada" del tratamiento en los artículos 13(2)(f), 14(2)(g) y 15(1)(h) RGPD[662];otras que puede interpretarse como integrado en el derecho a impugnar la decisión en el artículo 22(3) RGPD[663]; y también que no hay lugar a un derecho a una explicación en el RGPD[664].

Sin embargo, lo relevante en esta discusión no parece tanto el hecho de que el Reglamento incorpore un derecho a una explicación como tal, con dicha denominación, sino la extensión y alcance de los derechos efectivamente reconocidos por el RGPD. Por ello, a partir del debate generado en torno a esta cuestión, entraremos a continuación a valorar exclusivamente qué extensión y alcance tienen los derechos de información y acceso reconocidos en los artículos 13(2)(f), 14(2)(g) y 15(1)(h) RGPD, específicamente regulados para la toma de decisiones basada únicamente en el tratamiento automatizada y aplicable, *al menos*, a dicha clase de tratamiento. Y dentro de esa extensión y alcance, cabe preguntarse si lo que se ha venido definiendo por un derecho a una explicación tiene o no cabida. Empecemos por esta última cuestión.

Se considera que estos artículos del Reglamento engloban dos derechos distintos; por un lado, el derecho a conocer la existencia en sí de la toma de decisiones automatizada[665] y, por otro, un derecho a recibir información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas de dicho tratamiento para el interesado[666].

En este contexto, el derecho a una explicación ha sido generalmente concebido como el derecho a conocer las razones por las que se genera un determinado resultado sobre el que se basa la decisión que afecta de forma significativa a la persona interesada. Para comprender mejor qué implica esta clase de explicación es muy útil la clasificación utilizada por Wachter, Mittelstadt y Floridi, y que distinguen entre explicaciones que se

[662] Selbst y Powles, «Meaningful information and the right to explanation».

[663] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -».

[664] Wachter, Mittelstadt, y Floridi, «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation».

[665] Es decir, un derecho a levantar ese primer nivel de opacidad al que nos hemos referido en el marco teórico, en el que las personas interesadas no conocen siquiera que están siendo objeto de decisiones automatizadas. Vid. Apartado 4. Opacidad en la toma de decisiones automatizada basada en la elaboración de perfiles, en Capítulo 1. Marco teórico de la toma de decisiones automatizada basada en la elaboración de perfiles.

[666] Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 256.

refieren a la *funcionalidad del sistema*, es decir, la lógica, el significado, las consecuencias previstas y la funcionalidad general de un sistema automatizado de toma de decisiones; o a *decisiones específicas*, es decir, los fundamentos, las razones y las circunstancias individuales de una decisión automatizada específica[667]; a su vez, dichas explicaciones pueden producirse *ex ante*, de forma previa a la toma de una decisión particular[668]; o *ex post*, una vez que la decisión particular ha sido adoptada[669].

No parece haber discrepancias a la hora de considerar que los derechos a obtener información significativa sobre la lógica aplicada engloban explicaciones sobre la funcionalidad del sistema de forma ex ante o ex post. Sí, en cambio, encontramos discrepancias sobre si alcanza a explicaciones sobre decisiones específicas con carácter ex post.

Wachter et al. mantienen una posición contraria, esto es, entienden que el RGPD solo exige informar sobre funcionalidades del sistema de forma ex ante, y es que el criterio del CEPD parece ser el mismo: «*El artículo 15, apartado 1, letra h), establece que el responsable del tratamiento debe facilitar al interesado información sobre las consecuencias previstas del tratamiento, en vez de una explicación sobre una decisión «particular"*»[670].

Sin embargo, la doctrina mayoritaria considera sobre distintos argumentos que el derecho de acceso en el artículo 15(1)(h) establece que los responsables del tratamiento están obligados a revelar información sobre la lógica realmente empleada en la decisión

[667] Selbst y Powles discuten este marco para el debate, dado que los sistemas de aprendizaje automático son fundamentalmente deterministas y, por ende, predecibles. Por ello, para muchos sistemas una explicación completa a nivel de funcionalidad del mismo lo dirá todo sobre casos específicos, siendo posible a su vez generar una explicación sobre las decisiones específicas dados los datos de entrada de la persona interesada de los que ya dispone el responsable del tratamiento. Vid. Selbst y Powles, «Meaningful information and the right to explanation», 239.

[668] Como es obvio, con carácter ex ante únicamente pueden aportarse explicaciones sobre la funcionalidad del sistema.

[669] Wachter, Mittelstadt, y Floridi, «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation», 78.

[670] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 30. El problema aquí está en que el GT29 sostiene que sí existe un derecho a una explicación, pero que es en el propio Considerando 71 donde se introduce como salvaguarda del artículo 22(3). Interpretación, como ya se ha señalado, contraria a la jurisprudencia del TJUE puesto que los Considerando no tienen fuerza vinculante por sí mismos. Sobre este y otras confusiones introducidas por el GT29 en este aspecto, vid. Veale y Edwards, «Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling», 399-400.

particular, y no sólo sobre la funcionalidad general del sistema de una toma de decisiones algorítmica[671]. En particular, rebatiendo la idea de que la necesidad de informar sobre las consecuencias "previstas" conlleve que dicha información deba tener carácter ex ante.

De hecho, este hilo argumentativo que sostienen Wachter et al., podría llevarnos a considerar que el derecho a conocer la "existencia" de decisiones automatizadas, no engloba el derecho a conocer en particular qué decisiones se han adoptado sobre el interesado, sino simplemente a conocer si las mismas existen o no. Una posición poco compatible con el principio de transparencia.

Además, la transparencia algorítmica, jurídicamente hablando y de acuerdo con su carácter instrumental, debería abarcar la transparencia del proceso de toma de decisiones algorítmicas en la medida en que sea necesario para garantizar el respeto de otros derechos en virtud del RGPD. Por ello, se ha vinculado la necesidad de interpretar la inclusión de una explicación ex post sobre la decisión específica adoptada en los derechos de información y acceso, con la posibilidad de ejercer el derecho a impugnar dicha decisión[672]. Esto sería inviable a partir de informaciones genéricas sobre la existencia de decisiones automatizadas y sobre la funcionalidad del sistema. No solo eso, el hecho de que la información deba tener un carácter “significativo” parece también descartar que, una vez una decisión haya sido adoptada, pueda ser significativa una información aplicable a cualquier clase de dato de entrada -y no en particular a los que han servido como base de la toma de decisiones-[673].

Reforzando esta posición, parte de la doctrina se ha centrado en la interpretación de dicho carácter “significativo” que debe adoptar la información facilitada por el responsable del tratamiento. Malgieri y Comandé defienden la existencia de un derecho a la legibilidad

[671] Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 256; también en Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 114; Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 93.

[672] Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 114-17. En un sentido similar, Roig apunta a que solo puede impugnarse si se comprende cómo se ha tomado la decisión particular y sobre qué base, construyendo una analogía con el fundamento del artículo 24 CE y la indefensión: *nadie puede articular una defensa de sus intereses si no conoce los argumentos que se le oponen. (…) difícilmente podrán impugnarse clasificaciones incorrectas, evaluaciones basadas en previsiones imprecisas o que afecten negativamente a las personas, sin entender cómo se ha tomado la decisión y sobre qué base*. Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 59.

[673] Selbst y Powles, «Meaningful information and the right to explanation», 237.

del tratamiento automatizado a partir una interpretación gramatical del término "*meaningful*" en una doble vertiente: (a) "*meaningful*" en el sentido de "relevante, importante" garantiza la transparencia[674], (b) mientras que "*meaningful*" en el sentido de "destinado a mostrar el significado" garantiza la comprensibilidad o legibilidad de la toma de decisiones automatizada[675]. Para Selbst y Powles el énfasis de los artículos 5 y 12 RGPD en los principios transparencia y responsabilidad refleja una perspectiva normativa nueva y mejorada sobre el perfilado algorítmico y la toma de decisiones, que debe traducirse en una información útil, inteligible y accionable –respecto del resto de derechos reconocidos– para la persona interesada[676].

La propuesta de *test de legibilidad* de Malgieri y Comandé es un buen ejemplo de la clase de preguntas/respuestas que una información útil, relevante y comprensible debería motivar sobre la lógica aplicada de una decisión algorítmica, su importancia y las consecuencias previstas[677]. Muchas de ellas, sobre el propio origen del programa informático (¿ha sido validado de forma externa? ¿es accesible dicha validación?) o el uso del mismo (¿qué datos de entrada se utilizan y cuál es su origen y calificación normativa?), es probable que no cambien independientemente de si una decisión particular ha sido o no adoptada; mientras que otros, sobre todo referidos a los datos de salida (¿qué criterios han influido de forma determinante en la producción del resultado?)

[674] Desde mi punto de vista, respecto de la primera vertiente, lo que resulta importante o relevante se define también por el *timing* entre la adopción de una decisión y el deber de aportar la información. Es posible que al momento de aportar información por parte del responsable, conforme a los artículos 13 y 14, ya se haya adoptado la decisión automatizada – por su inmediatez – o se conozca con certeza qué datos de entrada se van a utilizar en el sistema, por ende, en dicho momento lo importante o relevante será ya la información sobre la decisión concreta que se va a adoptar y no la información sobre la funcionalidad del sistema. Ello con independencia de si estamos ante un deber de información de los arts. 13 y 14 o de un derecho de acceso del art. 15. Lo significativo es relativo desde una perspectiva temporal.

[675] Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 257. En el mismo sentido, Brkan sostiene que la información sobre la lógica de la decisión automatizada solo es significativa si la interesada puede *comprender* los factores y consideraciones en los que se basó la decisión, en Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 113. Vertiente que parece encontrar también su sentido en el artículo 12(1) RGDP: *El responsable del tratamiento tomará las medidas oportunas para facilitar al interesado toda información indicada en los artículos 13 y 14, así como cualquier comunicación con arreglo a los artículos 15 a 22 y 34 relativa al tratamiento, en forma concisa, transparente, inteligible y de fácil acceso, con un lenguaje claro y sencillo (…)*.

[676] Selbst y Powles, «Meaningful information and the right to explanation», 242. Aunque también se ha cuestionado ampliamente el alcance real de los derechos explícitamente reconocidos en el RGPD para contestar una decisión algorítmica, vid. Kluttz, Kohli, y Mulligan, «Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions», 147.

[677] Vid. Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 259-61.

o la implementación del sistema (¿qué consecuencias tiene sobre la situación jurídica o contractual del interesado?), variarán más en función de si se informa sobre un tratamiento realizado o uno previsible.

No obstante, una mejor técnica legislativa habría aportado soluciones más satisfactorias. La doctrina destaca en este sentido el derecho a una explicación introducido por *la Loi pour une République numérique*[678] y el artículo R311-3-1-2 en el *Code des relations entre le public et l'administration*[679] francés. Tal y como recogieron Edwards y Veale, esta ley estableció que en el caso de una decisión adoptada sobre la base de un tratamiento algorítmico -aplicándose también a sistemas automatizados de apoyo a la toma de decisiones-, las reglas que definen ese tratamiento y sus características principales deben ser comunicadas previa solicitud[680]. Derecho que fue posteriormente definido en el artículo R311-3-1-2 así: «*La Administración comunicará a la persona objeto de una decisión individual adoptada sobre la base de un tratamiento algorítmico, a petición de ésta, de forma inteligible y a condición de no vulnerar los secretos protegidos por la ley, la siguiente información: 1° El grado y el modo de contribución del tratamiento algorítmico a la decisión; 2° Los datos tratados y sus fuentes; 3° Los parámetros de tratamiento y, en su caso, su ponderación, aplicados a la situación del interesado; 4° Las operaciones realizadas por el tratamiento*»[681]. Desde luego, el margen de mejora en el texto del RGPD es considerable[682].

---

[678] *LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique*. Accesible aquí: https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829/

[679] Este artículo fue introducido por el *Décret n°2017-330 du 14 mars 2017* en el *Code des relations entre le public et l'administration*. Accesible aquí: https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000034195881

[680] Edwards y Veale, «Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?», 48.

[681] Traducción del original en francés: *L'administration communique à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes: 1° Le degré et le mode de contribution du traitement algorithmique à la prise de décision; 2° Les données traitées et leurs sources; 3° Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé; 4° Les opérations effectuées par le traitement*.

[682] En palabras de Roig sobre esta clase de derecho en la legislación francesa: *(…) así configurado, no sirve solamente al propósito de impugnar herramientas de decisión con sesgos discriminatorios; también permite recurrir ante la decisión que nos ha afectado, disponiendo de elementos para justificar las alegaciones. En este sentido, se enriquece el derecho a la intervención humana con una suerte de derecho previo a la resolución fundamentada y transparente. Ello debería evitar la intervención humana puramente formal, sin capacidad real de valorar la decisión o de rectificarla. Por consiguiente, la explicación basada en el sujeto sirve a varios propósitos: permite, en primer lugar, fundamentar la alegación del interesado; en segundo lugar, dota al humano que interviene posteriormente de los elementos necesarios para tomar*

2.3. Limitaciones de los derechos de información y acceso basados en el principio de transparencia

La discusión reflejada en el punto anterior parece estar llegando a cierto consenso sobre el alcance de estos derechos de información y acceso. El problema ahora está en hacer efectivos los mismos, dadas las limitaciones técnicas y normativas que se han encontrado[683]. Quizás, conscientes de ello, la doctrina ha decidido abandonar la discusión anterior.

Empecemos por el plano normativo. Como bien señala Palma Ortigosa, el RGPD reconoce que el derecho a la protección de datos personales no es un derecho absoluto, debiendo de lidiar con otros bienes e intereses jurídicos con los que puede entrar en juego[684]. En el articulado del Reglamento encontramos una única limitación expresa en este sentido, el artículo 15(4) RGPD dice que el derecho a obtener copia de los datos personales del tratamiento *no afectará negativamente a los derechos y libertades de otros.* Hay aquí dos cuestiones fundamentales a tratar: por un lado, la colisión entre los derechos de protección de datos de terceros y de la persona interesada, por otro, la colisión entre los secretos comerciales o la propiedad intelectual y los derechos de protección de datos de la persona interesada.

En cuanto a la colisión de derechos de protección de datos de distintas personas, hemos tenido ocasión de tratar esta cuestión al hablar de los perfiles o inferencias algorítmicas con la interpretación del TJUE en el caso *Nowak*. Si un mismo dato personal puede referirse a dos personas distintas, es igualmente posible que la información significativa sobre determinado tratamiento automatizado se corresponda con los datos personales de terceros. En este sentido, el responsable del tratamiento deberá realizar una ponderación adecuada de los derechos en conflicto, no obstante y conforme al Considerando 63: «*estas consideraciones no deben tener como resultado la negativa a prestar toda la información*

una eventual decisión correctora; y finalmente, si un órgano judicial llega a conocer el asunto, los términos de la disputa estarán igualmente más claros. Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 104.

683 En el marco teórico ya se ha hecho referencia a estas colisiones a partir de la definición de las distintas clases de opacidad. Vid. Apartado 4. Opacidad en la toma de decisiones automatizada basada en la elaboración de perfiles, en Capítulo 1. Marco teórico de la toma de decisiones automatizada basada en la elaboración de perfiles.

684 Palma Ortigosa, «Automated Decision-Making in the GDPR. Algorithms in the Scope of the Data Protection», 29.

al interesado». Y es que el propio RGPD aporta herramientas para que datos personales que se refieren a terceros cuyos derechos deban ser objeto de protección, puedan aportarse a la persona interesada como información significativa del tratamiento que le afecta; por ejemplo, la seudonimización definida en el artículo 4 RGPD[685].

Más discutida es la colisión entre derechos de terceros reconocidos en la normativa sobre secretos comerciales o propiedad intelectual, cuestión directamente abordada en el Considerando 63 RGPD, el cual dice que el derecho de acceso «*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*». Eso sí, recuerda el CEPD que, de acuerdo con el texto del mismo Considerando, 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[686].

Aunque este considerando -e incluso el articulo 15(4)- reconozca expresamente que los responsables del tratamiento, o terceros, tienen derechos e intereses relevantes en este contexto, no dejan igualmente de prever unos derechos de carácter preferente para los interesados, especialmente en relación con la información que debe proporcionar el responsable del tratamiento, por lo que no está justificado dar esa preferencia normativa a los secretos industriales sobre esta nueva legislación que regula y desarrolla derechos fundamentales[687].

Esta prevalencia se hace expresa al comparar el Considerando 35 de la Directiva de secretos comerciales[688] con el Considerando 63 del RGPD, mientras que este último "no debe afectar negativamente" los derechos sobre secretos comerciales o propiedad intelectual, el primero expresa que dicha normativa "no debe afectar" a los derechos y

[685] Art. 4(5) RGPD: «seudonimización»: el tratamiento de datos personales de manera tal que ya no puedan atribuirse a un interesado sin utilizar información adicional, siempre que dicha información adicional figure por separado y esté sujeta a medidas técnicas y organizativas destinadas a garantizar que los datos personales no se atribuyan a una persona física identificada o identificable.

[686] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 19.

[687] Selbst y Powles, «Meaningful information and the right to explanation», 242.

[688] Directiva (UE) 2016/943 de 8 de junio de 2016 relativa a la protección de los conocimientos técnicos y la información empresarial no divulgados (secretos comerciales) contra su obtención, utilización y revelación ilícitas.

obligaciones previstos en la normativa de protección de datos. La omisión de dicho calificativo ("negativamente") hace patente que la esfera de afectación en dicha colisión de intereses difiera, prevaleciendo la normativa de protección de datos[689].

A las limitaciones normativas que hemos señalado para los derechos de información y acceso, hay que sumar las limitaciones técnicas que se resumen a continuación[690]. Tal y como se ha explicado anteriormente, las funciones utilizadas por muchos de los algoritmos predictivos basados en aprendizaje automático para la toma de decisiones pueden ser demasiado complejas para que las comprendan los seres humanos, y en este sentido, no es interpretable para los humanos cómo el algoritmo ha llegado a una determinada decisión a partir de determinados datos de entrada. Como solución, se han propuesto por la literatura científica métodos de interpretabilidad que buscan, a fin de cuentas, simplificar la complejidad de dichos algoritmos.

La ventaja de estos métodos es que la mayoría no ponen en riesgo la revelación de información que comprometa los secretos comerciales o la propiedad industrial o intelectual[691]. Sin embargo, mientras que algunos de estos métodos pueden servir de apoyo para aportar información sobre la lógica aplicada en el tratamiento automatizado, no explican las previsibles consecuencias del mismo, ni justifican cómo se ha adoptado determinada decisión; por ello, Crabtee et al. concluyen que hay ciertas explicaciones que quedan necesariamente fuera de la toma de decisiones algorítmica y que no pueden ser proporcionadas por los métodos de interpretabilidad[692].

En el contexto de los derechos de información y acceso en el RGPD, se ha discutido especialmente el método de las explicaciones contrafactuales a partir de la investigación

[689] Vid. Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 262-64.

[690] También en el marco teórico nos hemos referido a esta clase de opacidad, calificada como "opacidad inherente al modelo". Vid. Apartado 4. Opacidad en la toma de decisiones automatizada basada en la elaboración de perfiles, en Capítulo 1. Marco teórico de la toma de decisiones automatizada basada en la elaboración de perfiles.

[691] Edwards y Veale, *Slave to the Algorithm? Why a Right to Explanationn is Probably Not the Remedy You are Looking for*, 18:26 y ss.

[692] Crabtree, Urquhart, y Chen, «Right to an Explanation Considered Harmful», 5; también en este sentido, Selbst y Barocas, «The Intuitive Appeal of Explainable Machines». Acerca de la viabilidad de los distintos métodos técnicos para la explicabilidad algorítmica, las distintas clases de explicaciones posibles y la compatibilidad de estas explicaciones y métodos con el RGPD, vid. Brkan y Bonnet, «Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas».

de Wachter, Mittelstadt y Russell. Frente a los métodos que tratan de esclarecer la lógica aplicada por el algoritmo, las explicaciones contrafactuales se centran en explicar las circunstancias -datos de entrada- que han llevado a determinado resultado, y cómo deberían variar para alcanzar un resultado más deseable para el interesado, de forma que es un método que no solo ayuda a comprender, sino también a actuar a partir de tal comprensión[693]. De este modo, las personas interesadas no necesitarían entender la lógica automatizada para extraer información significativa sobre la toma de decisiones[694].

Se discute si la– utilización de explicaciones contrafactuales podría satisfacer el cumplimiento del derecho a una explicación particular sobre la decisión adoptada -en los términos antes expuestos-, Roig expresa dudas al respecto, aunque destaca la ventaja de la sencillez de esta clase de explicaciones, a la vez que su utilidad para el usuario que pretendiese obtener una decisión alternativa en el futuro[695]. Este autor advierte, no obstante, que las explicaciones contrafactuales no son neutrales: «*(…) deben, al contrario, tomarse decisiones sobre qué atributos revelar y cuál es el orden de los valores del mundo real al cual se aplican. Mientras quien toma la decisión disponga de discrecionalidad y ambigüedad, ello le otorga sin duda un poder que puede utilizar para diferentes fines*»[696].

Es cierto que esta propuesta -al igual que otros métodos interpretativos- puede aportar información y explicaciones útiles y sencillas a la persona interesada, ahora, ello no puede

[693] Wachter, Mittelstadt, y Russell, «Counterfactual explanations without opening the black box: automated decisions and the GDPR VO - 31 RT - Journal Article», 843. Por ejemplo, la solicitud de préstamo de un individuo es rechazada por un sistema automatizado. Una explicación contrafactual proporcionaría al individuo una evaluación del cambio mínimo necesario para que la solicitud fuera aceptada. De modo que la explicación contrafactual diría que el préstamo se habría concedido si el solicitante solicitara un préstamo de 10.000 euros o menos siendo sus ingresos de entre 25.000 y 30.000 euros.

[694] Gacutan y Selvadurai, «A statutory right to explanation for decisions generated using artificial intelligence», 143. Este argumento denota, no obstante, un cierto paternalismo hacia las personas interesadas, al igual que cuando se afirma que este método aportaría una información más relevante que la revelación del código algorítmico puesto que las interesadas no tienen la formación para comprenderlo. Una cosa es que, efectivamente, este método pueda facilitar una información simplificada y útil, otra muy distinta es que dicha información sea más valiosa en sí que la información sobre la lógica aplicada o la misma revelación del algoritmo.

[695] Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 79.

[696] Roig, 81. También Barocas, Selbst y Raghavan expresan algunas de las limitaciones técnicas del modelo, dadas las condiciones necesarias para su correcto funcionamiento: (1) que los valores de los atributos se correspondan con acciones realmente posibles para el usuario (2) que pueda cuantificarse el valor alternativo a partir de únicamente los datos que han servido para entrenar al algoritmo (3) que los atributos indicados sean únicamente relevantes para la decisión discutida y no para otros ámbitos y (4) que el modelo sea estable, no cambie con el tiempo y admita además resultados de tipo binario. Vid. Barocas, Selbst, y Raghavan, «The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons».

suponer que el responsable del tratamiento de por supuesto con la aplicación de estos métodos, el cumplimiento de los derechos de información y acceso del RGPD[697]. Por ejemplo, si las explicaciones contrafactuales no satisfacen adecuadamente el acceso a la información significativa sobre la lógica aplicada por el algoritmo, el responsable del tratamiento deberá aportar -de forma adicional a las explicaciones contrafactuales- información particular sobre este aspecto[698].

Una última limitación parece estar a caballo entre lo técnico y lo normativo, el Considerando 63 expresa que, si el responsable del tratamiento trata una gran cantidad de información relativa al interesado, éste debe estar facultado para*:* «*solicitar que, antes de facilitarse la información, el interesado especifique la información o actividades de tratamiento a que se refiere la solicitud*».

Esta limitación fue aplicada en los casos resueltos por el Tribunal de Distrito de Ámsterdam sobre las disputas entre los *drivers* y las compañías Uber y Ola[699]. Dicho tribunal denegó la solicitud de acceso a varios de los parámetros, considerados datos personales o elaboración de perfiles, a través de los cuales Uber y Ola regulaban la actividad de los *drivers*. La denegación del acceso según el tribunal respondía a que no habían especificado con suficiente detalle la solicitud de acceso sobre la base del Considerando 63 RGPD[700].

---

[697] También en esta línea argumental, Kroll afirma que tratar estos sistemas como si fueran imprevisibles o incontrolables, en este caso inexplicables, ignoraría el hecho de que son artefactos humanos, construidos con un propósito por alguna agencia humana que debe ser responsable de sus comportamientos, en Kroll, «The fallacy of inscrutability», 7.

[698] Las explicaciones contrafactuales -apoyadas sobre los secretos industriales y propiedad intelectual- no pueden convertirse en el ejemplo paradigmático aportado por Korff para explicar la creciente incapacidad para impugnar la toma de decisiones automatizada: *(…) a company (or State agency) will tell you it will not give you a loan, or will not invite you to an interview (or has placed you on a terrorist "no-fly" list, or worse), "because the computer said so": because the computer generated a "score" based on a profile, that exceeded or did not reach some predetermined basic level. If you ask for an explanation (if, that is, you actually find out that such an automated decision has been made on you), the company or agency (or at least the person you are dealing with) is likely to be unable to explain the decision in any meaningful way. They might provide you with examples of some of the information used (age, income level, whatever), but they will not give you the underlying algorithm - partly because the respondent him- or herself does not know or understand that algorithm, which is in any case constantly dynamicaly changing, and partly because the algorithm is a "commercial secret"*. Korff, «Comments on Selected Topics in the Draft EU Data Protection Regulation», 25.

[699] Sentencias del Tribunal de Distrito de Ámsterdam [*Rechtbank Amsterdam*] de 11 de marzo de 2021; *Uber deactivation case* (ECLI:NL:RBAMS:2021:1018); *Ola transparency case* (ECLI:NL:RBAMS:2021:1019); y *Uber transparency request case* (ECLI:NL:RBAMS:2021:1020).

[700] Al respecto, vid. Lazcoz, «Automated decision-making under Amsterdam's District Court judgements: Drivers v. Uber and Ola».

Desde mi punto de vista, hay dos cuestiones que llaman la atención en esta resolución. La primera, la aplicación directa que se hace del texto del Considerando para denegar un derecho reconocido en el articulado del RGPD. La segunda, quizás más grave, que dicha denegación se produce con base en algunos parámetros sobre los que los demandantes habían manifestado que la información aportada por Uber y Ola inicialmente era totalmente incomprensible[701]. Es decir, el tribunal desplaza así una carga absolutamente desproporcionada sobre las personas interesadas en el ejercicio de sus derechos.

En definitiva, hay mucho trabajo jurídico e interdisciplinar por realizar en este ámbito. Por un lado, el alcance de los derechos de información y acceso en relación con la toma de decisiones automatizada debe definirse con mayor detalle[702]; la interpretación por los tribunales de los EEMM puede ser de ayuda, ahora, sería recomendable que el CEPD enmendase las Directrices del GT29 en este aspecto, ya que son especialmente erráticas. Del mismo modo, las limitaciones normativas a las que nos hemos referido -en particular los derechos relativos a secretos industriales y de propiedad intelectual-, no pueden convertirse en una excusa en manos de los responsables del tratamiento para limitar el derecho a la protección de datos de las personas interesadas. La limitación de un derecho fundamental debe ser la excepción, debidamente justificada y ponderada, y no la regla general. En cuanto a la opacidad inherente a los complejos modelos algorítmicos, es una cuestión que indudablemente limita la interpretabilidad y explicabilidad de los mismos. Y tiene cierta relevancia normativa, pero el cumplimiento del Reglamento no puede hacerse depender del desarrollo de métodos más o menos satisfactorios a nivel técnico.

El RGPD impone que el responsable del tratamiento haga un uso responsable, transparente, leal y exacto, entre otros, de los datos personales con independencia de que los trate con la tecnología más rudimentaria o la más avanzada -más si cabe, en tal caso-. Por ello, ha de ser capaz de explicar las lógicas a las que obedece el tratamiento automatizado en su conjunto que, libremente, decide aplicar sobre los datos personales de los interesados, y no ampararse en la opacidad algorítmica para dejar de informar sobre

[701] Sentencias del Tribunal de Distrito de Ámsterdam [*Rechtbank Amsterdam*] de 11 de marzo de 2021; *Ola transparency case*, par. 4.26; y *Uber transparency request case*, par. 4.55-4.56.

[702] Más contundente, de Vries pone el foco sobre la normativa y concluye que, dado el modo en que la protección de datos pasa por alto varios aspectos clave de los algoritmos de aprendizaje automático, en particular el modo específico de producción de la *verdad* en estos algoritmos, hasta la fecha: *no informational right that has ever encountered a profiling algorithm as a profiling algorithm*. de Vries, «Machine learning/Informational fundamental rights: Makings of sameness and difference», 418.

las determinantes operaciones de tratamiento sobre las que tiene -o debería tener- pleno control[703].

3. Derecho a impugnar las decisiones automatizadas

¿Hay en el RGPD un derecho a rebatir, refutar o discutir las inferencias algorítmicas y decisiones automatizadas que nos afectan?

Desde mi punto de vista, el alcance de este derecho -conjunto de derechos- como tercer pilar en la regulación de la toma de decisiones automatizada, el nivel de permeabilidad sobre las decisiones automatizadas y elaboración de perfiles es lo que determinará la -tímida- consecución de uno de los mayores retos a los que se enfrenta la privacidad según Turégano, esto es, convertir el uso de las tecnologías de la información en un espacio público igualitario, plural y abierto[704].

Al referirnos a los fundamentos de la regulación de la toma de decisiones automatizada introducida en el artículo 15 de la DPD, se ha señalado que la Comisión en los trabajos preparatorios aludía al temor kafkiano relacionado con la pérdida de control sobre las decisiones que afectan a uno como interesado: «*Con esta disposición se pretende proteger el interés del interesado en participar en aquellas decisiones que sean importantes para él. El uso de perfiles detallados basados en datos personales por parte de importantes instituciones públicas y privadas priva al interesado de la posibilidad de influir en los procesos decisorios de dichas instituciones cuando esas decisiones se toman únicamente sobre la base de su perfil personal*»[705]. El creciente uso de las técnicas de elaboración de perfiles, que provocaba que las personas tuvieran cada vez menos control y capacidad de influencia sobre las decisiones que les afectaban, fue el primer catalizador para la inclusión de ese artículo 15 DPD[706].

---

[703] En otras palabras, el considerando 15 del RGPD: A fin de evitar que haya un grave riesgo de elusión, la protección de las personas físicas debe ser tecnológicamente neutra y no debe depender de las técnicas utilizadas.

[704] Turégano Mansilla, «Los valores detrás de la privacidad», 279.

[705] Comisión de las Comunidades Europeas, Comunicación de la Comisión «sobre la protección de las personas en lo referente al tratamiento de datos personales en la Comunidad y a la seguridad de los sistemas de Información». Bruselas 24.09.1990. COM(90) 314 final - SYN 288, p. 22.

[706] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 83. Vid. también esta problemática en los términos expuestos por Korff, "la creciente incapacidad para la impugnabilidad de los perfiles", en Korff, «Comments on Selected Topics in the Draft EU Data Protection Regulation», 25 y ss.

En este sentido, hemos argumentado que la intervención humana, *out of the loop*, como salvaguarda bajo requerimiento del interesado en el artículo 22(3) RGPD respondía a este fundamento. Por ello, la doctrina señala que dicha intervención tiene por objeto solicitar una segunda resolución, una revisión, en la que un agente humano puede tener en cuenta también el punto de vista del interesado[707], cuestión que se reafirma en las directrices del GT29[708]. Así, concluía más arriba que la "significancia" de esta intervención humana estaría directamente vinculada con la posibilidad del interesado de ejercer sus derechos a expresar su punto de vista y a impugnar la decisión.

Del mismo modo, Roig insiste en el aspecto instrumental de la transparencia, no se trata únicamente de estar informado, sino de disponer de posibilidades de ejercer los demás derechos reconocidos en el RGPD[709]. Es decir, entre los derechos que se reconocen en el RGPD, algunos de ellos posibilitan el ejercicio de control e influencia sobre el tratamiento de datos personales que afecta al interesado, ahora bien, ¿cuáles son esos derechos para ejercer ese control e influencia directa y qué alcance tienen?

A continuación, vamos a analizar los mecanismos de impugnación ex post que permiten ejercer dicho control por parte de los interesados, divididos de la misma forma que en los análisis de los pilares anteriores; por un lado, los mecanismos disponibles sobre las inferencias o perfiles que el responsable genera y, por otro, los mecanismos sobre la toma de decisiones automatizada en base únicamente a los mismos. En opinión de Wachter y Mittelstandt, dichos mecanismos deberían fortalecer el diálogo entre el responsable del tratamiento y la persona interesada cuando ésta cuestiona la exactitud o razonabilidad de una inferencia o de la decisión tomada en base a la misma, sin que ello suponga la imposición del criterio de la persona interesada[710]. Por mi parte, añadiría, deben fortalecer

[707] Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems»; Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations».

[708] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 30.

[709] Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 47. No olvidemos que los derechos de transparencia pueden servir igualmente para el ejercicio de derechos que no necesariamente estén recogidos en el RGPD, permitiendo, por tanto, acudir a otras bases normativas para la impugnación de las decisiones que correspondan. Cuestión que en el ámbito sectorial puede adquirir mucha más relevancia que los mecanismos de impugnación reconocidos en el RGPD y que se analizan a continuación.

[710] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI», 588-89.

un diálogo sobre los principios y derechos del RGPD, que tenga como resultado la imposición del cumplimiento del mismo.

3.1. Derecho a rectificar las inferencias algorítmicas: ¿cuál es el alcance del principio de exactitud respecto de la elaboración de perfiles?

En el RGPD no se reconoce un derecho a impugnar las inferencias algorítmicas o perfiles como tal, dado que este derecho está reconocido exclusivamente como salvaguarda para las decisiones habilitadas por el artículo 22(2) -basadas únicamente en el tratamiento automatizado-. Ello no impide que el interesado no disponga de herramientas para impugnar cualquier clase de inferencia que se realice sobre él y que, como ya hemos argumentado anteriormente, son también sus datos personales.

La rectificación y supresión de datos personales son operaciones derivadas directamente del principio de exactitud, y el RGPD obliga al responsable del tratamiento a aplicar las mismas sobre los datos personales que sean inexactos con respecto a los fines para los que se tratan -art. 5(1)(d)-[711], pero también faculta a las personas interesadas a exigir el cumplimiento de este principio a través de los derechos de rectificación y supresión -arts. 16 y 17-.

Desde un punto de vista gramatical rectificar es reducir algo a la exactitud que debe tener[712], la rectificación como tal la realiza en todo caso el responsable del tratamiento y, por ende, ese derecho de rectificación es en realidad un derecho del interesado a impugnar la (in)exactitud de los datos personales. Del mismo modo, el derecho a impugnar las decisiones automatizadas en el artículo 22, permite rebatir una inferencia -que produce determinados efectos basada únicamente en el tratamiento automatizado- y solicitar su

[711] Bajo esta perspectiva el principio de exactitud no depende del contacto con la persona interesada, vid. Hallinan y Zuiderveen Borgesius, «Opinions can be incorrect (in our opinion)! On data protection law's accuracy principle», 3. Las Directrices del GT29 amplían sobre las obligaciones del responsable en la elaboración de perfiles, advirtiendo del riesgo de no tener en cuenta la exactitud de los datos personales en todas las fases del perfilado (recogida y análisis de datos, creación y aplicación de los perfiles): *Si los datos utilizados en un proceso de decisión automatizada o de elaboración de perfiles son inexactos, cualquier decisión o perfil resultante será defectuoso. Las decisiones pueden tomarse sobre la base de datos desactualizados o de la incorrecta interpretación de datos externos. Las inexactitudes pueden provocar predicciones o afirmaciones inadecuadas sobre, por ejemplo, la salud, el crédito o el riesgo de seguro de una persona*. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 13.

[712] Primera acepción del término "rectificar" en el Diccionario de la lengua española de la Real Academia Española.

"rectificación" al responsable del tratamiento[713]. En este sentido, en el contexto de las inferencias algorítmicas puede considerarse que el derecho de rectificación -art. 16- y el derecho de impugnar las decisiones automatizadas -art. 22(3)- son derechos hermanos, que comparten un mismo objetivo: impugnar ante el responsable determinada inferencia algorítmica[714] y solicitarle la rectificación del tratamiento de datos personales.

Centrados en el derecho de rectificación, la doctrina lo considera un derecho de utilidad en el contexto algorítmico, que permite impugnar no sólo el dato inexacto, sino cualquier categoría o grupo al que se haya designado al interesado a través de la elaboración de perfiles[715]. En las Directrices adoptadas por el CEPD encontramos varios puntos clave que nos hablan de su utilidad: (1) las inferencias predictivas aumentan el riesgo de inexactitud, lo cual resalta el valor y la necesidad de este derecho de rectificación; (2) la necesidad de utilizar datos exactos en toda la fase de la elaboración de perfiles refuerza la importancia de ofrecer una información clara sobre los datos personales que se procesan, resaltando de nuevo el valor instrumental de la transparencia y su conexión con la posibilidad de control de los interesados sobre sus datos personales; y (3) los derechos de rectificación y de supresión se aplican tanto a los datos de entrada como a los datos de salida, lo cual refuerza la idea de control sobre todas las fases de la elaboración de perfiles [716].

En su artículo sobre un derecho a unas inferencias razonables, Wachter y Mittelstandt argumentan que el actual artículo 16 RGPD ofrece un mecanismo válido para impugnar inferencias no razonables. En su artículo, distinguen entre inferencias verificables (nombre, edad, ingresos, estado civil) y no-verificables (opiniones subjetivas), y entienden que el derecho de rectificación se basa implícitamente en la noción de verificación, lo que significa que se puede demostrar que un dato no es válido (es decir,

[713] No obstante, parece que la rectificación en el ámbito del derecho a impugnar una decisión automatizada puede abarcar fundamentos más amplios que el derecho de rectificación, que está constreñido a la aplicación del principio de exactitud, mientras que las decisiones automatizadas parecen abarcar la posibilidad de impugnar y rectificar sobre fundamentos más amplios como el principio de lealtad, dado el mandato de no discriminación contenido en el Considerando 71.

[714] En caso del artículo 16 RGPD las inferencias serán cualquiera que tenga carácter personal, mientras que para el artículo 22(3) solo aquéllas que produzcan determinados efectos basados únicamente en el tratamiento automatizado.

[715] Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 56.

[716] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 13 y 19.

es inexacto o incompleto) y, por lo tanto, el responsable debe corregirlo[717]. No obstante, señalan, las predicciones probabilísticas de los algoritmos son habitualmente no verificables, o bien por ser *inherentemente subjetivas* (prestatario de "alto riesgo")[718] o puramente predictivas (el interesado solicitará un crédito en los próximos dos años)[719]. Por último, reiteran la idea de que para la jurisprudencia del TJUE el cometido de la normativa de protección de datos no es evaluar la exactitud del razonamiento que subyace a las decisiones y evaluaciones, ni la exactitud de las propias decisiones y evaluaciones -para ello habría que acudir a normativas sectoriales destinadas al efecto-[720].

Hay varios puntos a considerar aquí.

En primer lugar, sobre si la normativa de protección de datos puede evaluar la exactitud del razonamiento que subyace a las evaluaciones, hay una diferencia fundamental que se está pasando por alto: la elaboración de perfiles es un tratamiento automatizado de datos personales sometido a las reglas de la normativa de protección de datos. En este sentido, la inferencia realizada por el evaluador en el caso *Nowak* al estimar el nivel de acierto en determinada pregunta de un examen, es una actividad cognitiva fuera del ámbito de regulación del tratamiento de datos personales. Sin embargo, el tratamiento de datos automatizado que realiza una inferencia estadística sobre determinado individuo es una actividad regulada por el RGPD y sometida a las normas y principios dictadas por el mismo.

Esta cuestión suscita preguntas muy relevantes sobre las limitaciones de la distinción verificable y no verificable y la aplicación del derecho de rectificación, ¿pueden equipararse las inferencias no verificables basadas en un criterio fundamentalmente

[717] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI», 548.

[718] No comparto que esto se trate de una inferencia inherentemente subjetiva, o al menos equiparable a una opinión subjetiva en términos normativos.

[719] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI», 548-49. Hallinan y Zuiderveen hacen una distinción ulterior entre los datos no verificables, dentro de lo que consideran "opiniones", y son por un lado los "hechos" que constituyen una opinión y, por otro, los "marcos interpretativos" de dichos hechos, pudiendo ser las opiniones puramente humanas, algorítmicas o mixtas. Definen opinión así: *una afirmación sobre una entidad, construida a partir de hechos sobre esa entidad sometidos a algún marco interpretativo para producir nuevos hechos probables sobre la entidad*. Hallinan y Zuiderveen Borgesius, «Opinions can be incorrect (in our opinion)! On data protection law's accuracy principle», 6.

[720] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI», 539-40.

estadístico a una inferencia basada en criterios subjetivos? ¿Cuál es el margen de autonomía que determina el RGPD para establecer inferencias algorítmicas no verificables sobre terceros? Y, por ende, ¿cuál es el alcance del derecho de rectificación sobre las mismas?

Tratando de alcanzar algunas respuestas más satisfactorias en el contexto de las inferencias algorítmicas, pueden distinguirse dos clases de rectificaciones:

Primero, rectificaciones sobre la misma base estadística utilizada por el responsable del tratamiento, aportando la interesada nuevos datos o actualización de los datos de entrada utilizados por el responsable del tratamiento. Esta clase de rectificación no parece ofrecer dificultad alguna a la hora de aplicar el artículo 16 del RGPD, sea la inferencia algorítmica verificable (la persona interesada tiene un inmueble) o no (la interesada adquirirá un inmueble en los próximos dos años). El derecho de rectificación no solo se aplica sobre los datos de entrada y salida utilizados por el responsable en sí, sino que también prevé el derecho del interesado a complementar los datos personales con información adicional[721], conforme al texto del mismo artículo 16 RGPD: «*Teniendo en cuenta los fines del tratamiento, el interesado tendrá derecho a que se completen los datos personales que sean incompletos, inclusive mediante una declaración adicional*».

Este punto, además de revelar una conexión con el derecho a expresar su punto de vista para las decisiones automatizadas en el artículo 22(3) RGPD, pone de manifiesto que, con independencia de que el responsable lleve a cabo el perfilado bajo uno u otro método estadístico y de si el resultado es verificable o no, la persona interesada tiene derecho a que sus datos sean tratados de forma completa y fiable conforme a dichos métodos decididos y aplicados por el responsable[722].

De hecho, esta interpretación es plenamente compatible con lo resuelto en el caso *Nowak*, donde el TJUE determinó que sí podían considerarse inexactas las anotaciones de un examinador si la evaluación no tenía en cuenta todas las respuestas o si, en definitiva, las anotaciones no se correspondían con los datos que debía valorar -en base a criterios establecidos por el propio responsable del tratamiento, una vez más-: «*En cambio, es posible que se den situaciones en las que las respuestas de un aspirante en un examen y*

[721] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 19.

[722] Podría decirse que el perfilado debe regirse por una igualdad formal en este sentido dentro del RGPD.

las correspondientes anotaciones del examinador sean inexactas, en el sentido del artículo 6, apartado 1, letra d), de la Directiva 95/46; por ejemplo, cuando por error las hojas de los exámenes se hayan entremezclado de tal modo que las respuestas de otro aspirante se hayan atribuido al aspirante afectado, o cuando se haya perdido una parte de los folios que contienen las respuestas de ese aspirante, dando lugar a que esas respuestas queden incompletas, o incluso cuando las eventuales anotaciones del examinador no documenten correctamente la valoración que este ha dado a las respuestas del aspirante de que se trate»[723].

Si esta clase de rectificación cabe sobre las anotaciones subjetivas, con más razón ha de reconocerse un derecho de rectificación sobre el proceso de elaboración de perfiles en el conjunto del tratamiento de datos, de acuerdo a la interpretación asumida por el CEPD, en base a las propias reglas estadísticas establecidas por el responsable para dicha actividad de perfilado. Esta clase de rectificación incide sobre la exactitud de los datos -de entrada- que son tomados como fundamento para la creación de la inferencia y su rectificación puede, o no, conllevar a su vez una rectificación de los datos de salida -el resultado podría llegar a ser el mismo, a pesar de rectificar los datos de entrada-.

El segundo supuesto es la rectificación del propio método estadístico para alcanzar la inferencia, o si se quiere en los términos propuestos por Hallinan y Zuiderveen, rectificación del marco interpretativo propuesto para la generación de opiniones algorítmicas[724]. Para analizar el alcance de la aplicación del principio de exactitud y del derecho de rectificación en estos marcos interpretativos sobre los que se generan las inferencias, hay que resaltar esa diferencia fundamental arriba indicada con las anotaciones u opiniones subjetivas fruto de procesos cognitivos humanos. El tratamiento automatizado inferencial en sí entra en el ámbito de aplicación del RGPD.

[723] STJUE de 20 de diciembre de 2017, caso *Peter Nowak contra Data Protection Commissioner* (Asunto C-434/16), par. 54.

[724] Parece no haber dificultades en acercar los términos empleados en su artículo: *Así pues, los marcos interpretativos pueden considerarse en función de la precisión probable con la que la información que producen refleje la realidad. Así, en términos de opiniones sobre datos personales, se confiará en que ciertos marcos interpretativos proporcionen datos personales más fiables y precisos que otros*. Hallinan y Zuiderveen Borgesius, «Opinions can be incorrect (in our opinion)! On data protection law's accuracy principle», 8. Aunque utilicen como base los "marcos interpretativos" para poder hablar de opiniones humanas/algorítmicas, al aplicar éstos a las opiniones algorítmicos parece obvio que nos estamos refiriendo a los métodos estadísticos.

Ello nos llevaría a la necesaria aplicación del principio de exactitud sobre este tratamiento. El GT29 parece reconocer este extremo en varias ocasiones, cuando expone que las decisiones pueden no solo tomarse sobre la base de datos desactualizados, sino también de la *incorrecta interpretación* de datos externos y que incluso si los datos se registran de forma precisa: «*el conjunto de datos puede no ser plenamente representativo o los análisis pueden contener desviaciones ocultas*»[725]. Más tarde, sobre el derecho de rectificación, dice que los datos de entrada pueden ser inexactos o irrelevantes, pero también puede haber algún error con el algoritmo utilizado para identificar correlaciones[726]. Estas alusiones parecen estar apuntando claramente a la aplicación del principio de exactitud sobre el método estadístico o el marco interpretativo que sirve como fundamento para la creación de inferencias.

En este sentido, Hallinan y Zuiderveen señalan que, dado que algunos marcos interpretativos producen opiniones más precisas que otros, en el plano normativo pueden trazarse líneas entre los marcos interpretativos capaces de producir información "suficientemente precisa" y los que son incapaces. Y concluyen que el RGPD encuentra legitimidad a la hora de trazar esa línea a través de la consideración de la razón fundamental que subyace al principio de exactitud: «*proteger a los individuos de ser objeto de tergiversaciones, y de las consecuencias de las mismas, a través de sus datos personales*»[727].

En definitiva, parece razonable defender la aplicación del principio de exactitud, y del derecho de rectificación, sobre estos marcos interpretativos o métodos estadísticos que configuran inferencias personales.

No obstante, más complicado parece, primero, determinar esa línea que el RGPD define sobre la precisión exigible al responsable del tratamiento -y que servirá de base para la solicitud de rectificación de la inferencia que afecta al interesado- y, segundo, llevar a la práctica este derecho individual.

[725] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 13.

[726] Grupo de Trabajo sobre Protección de Datos del Artículo 29, 19.

[727] Hallinan y Zuiderveen Borgesius, «Opinions can be incorrect (in our opinion)! On data protection law's accuracy principle», 10.

Sobre el segundo aspecto me remito a las -arriba analizadas- limitaciones técnicas y normativas de los derechos de información y acceso sobre las lógicas aplicadas por los algoritmos. Sobre cuál es la exactitud exigible en este contexto por el RGPD, la jurisprudencia parece establecer que no hay un enfoque único para toda clase de tratamiento, sino que será el contexto de la finalidad del tratamiento de los datos personales lo que determine la exactitud exigible. Así expresado en el caso *Nowak* por el TJUE: «(…) *el carácter exacto y completo de los datos personales debe ser apreciado atendiendo a los fines para los que fueron recabados*»[728]. Ello implica también que dependiendo de la aplicación sectorial en la que devenga esta clase de tratamiento -en los que pueden coexistir normas que se refieran ya a la exactitud que deben adoptar los marcos interpretativos utilizados-, los estándares normativos para definir dicha exactitud contextual podrían modularse[729].

Por último, cabe preguntarse si el derecho de rectificación permite impugnar inferencias algorítmicas sobre fundamentos distintos del principio de exactitud, pongamos por razones de licitud o lealtad.

Más claro parece que los principios del Reglamento deban dar lugar a rectificación en el marco del derecho a impugnar la decisión del 22(3), dado que el considerando 71 puede considerarse un mandato directo antidiscriminatorio en relación con las medidas de salvaguarda de dicho apartado en el artículo 22 RGPD[730]. Es decir, la corrección de inexactitudes y de errores parecen vincularse aquí directamente con efectos discriminatorios y con la aplicación del principio de lealtad, frente al carácter más formal que, aparentemente, guarda el derecho de rectificación en aplicación del principio de exactitud en el artículo 16.

---

[728] STJUE de 20 de diciembre de 2017, caso Peter Nowak contra Data Protection Commissioner (Asunto C-434/16), par. 53.

[729] Así lo sugieren expresamente Hallinan y Zuiderveen para el tratamiento de datos en el contexto diagnóstico, en Hallinan y Zuiderveen Borgesius, «Opinions can be incorrect (in our opinion)! On data protection law's accuracy principle», 9.

[730] Zuiderveen Borgesius, «Strengthening legal protection against discrimination by algorithms and artificial intelligence», 1579. Considerando 71 RGPD: *(…) aplicar medidas técnicas y organizativas apropiadas para garantizar, en particular, que se corrigen 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*.

No obstante, el considerando 72 RGPD establece claramente que la elaboración de perfiles: «*está sujeta a las normas del presente Reglamento que rigen el tratamiento de datos personales, como los fundamentos jurídicos del tratamiento o los principios de la protección de datos*». Por dicha razón, entiendo que el derecho de rectificación respecto de las inferencias algorítmicas debe interpretarse ampliamente junto con este considerando no vinculante, permitiendo la rectificación de inferencias que resulten también contrarias a los principios de licitud o lealtad. En cualquier caso, el propio considerando 72 dispone que el CEPD debe tener la posibilidad de formular orientaciones en este contexto sobre la elaboración de perfiles, lo cual resulta más que recomendable para este propósito.

3.2. Derecho a impugnar las decisiones basadas únicamente en el tratamiento automatizado

Entre las medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado para la toma de decisiones basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar, el artículo 22(3) RGPD reconoce, en última instancia, el derecho del interesado a impugnar la decisión. Sorprende de las Directrices del GT29 que el apartado dedicado a las medidas de salvaguarda del 22(3) es extremadamente conciso y apenas desarrolla los derechos de la persona interesada a obtener intervención humana por parte del responsable, a expresar su punto de vista y a impugnar la decisión, más allá de recalcar que éstas son las medidas mínimas a incluir por el responsable[731]. Lo cual amplía más si cabe el margen interpretativo que el texto deja sobre estas medidas[732].

Voces autorizadas en la doctrina han considerado que este derecho es la *columna vertebral* del conjunto de medidas de salvaguarda concebidas en la disposición kafkiana, es más, el fundamento del resto de medidas, tanto el derecho a expresar su punto de

[731] Vid. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 30-31. Como señalo en el apartado sobre los derechos de información y acceso, el GT29 añade también aquí el derecho a obtener una explicación entre estas medidas de salvaguarda, tal y como se recoge en el Considerando 71, a pesar de que se omite por el 22(3) RGPD.

[732] Vid. el considerable esfuerzo de Geburczyk al abordar el solapamiento normativo de los procedimientos administrativos nacionales y el RGPD, en Geburczyk, «Automated administrative decision-making under the influence of the GDPR – Early reflections and upcoming challenges».

vista[733] como el derecho a la intervención humana[734], han sido vinculadas de forma instrumental con la impugnación de la decisión automatizada, como condición necesaria para poder ejercitar dicho derecho. A su vez, entre las cuestiones que se han abordado en la doctrina sobre el derecho a impugnar la decisión, destaca el alcance también instrumental de los derechos de información y acceso para poder ejercer este derecho[735]. Así, se ha afirmado que los derechos de acceso e información relativos a las decisiones automatizadas sólo pueden aplicarse de forma efectiva si contribuyen a los derechos al "debido proceso" previstos en el artículo 22(3) RGPD[736].

No es extraño que este conjunto de medidas de salvaguarda haya sido catalogado como un derecho al debido proceso en la toma de decisiones automatizada puesto que tienen evidentes similitudes con derechos de origen procesal. Es más, en el análogo artículo 11 de la Directiva 680/2016[737], los derechos a impugnar su punto de vista y a impugnar la

[733] El derecho a expresar su punto de vista es un requerimiento procedimental más que un derecho en sí mismo, sin el derecho a impugnar la decisión no tendría el mismo valor. Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 58. De esta forma, las personas interesadas no solo tienen derecho a recurrir para obtener una nueva decisión, sino que también pueden aportar información que pueda ser relevante para reconsiderar el resultado inicial. Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations», 11.

[734] En relación a los fundamentos de la intervención humana *out of the loop* ya hemos tenido ocasión de resaltar que esta clase de intervención se fundamenta en la posibilidad de que el interesado influya sobre las decisiones que le afecten significativamente, de forma que la intervención humana se configura de nuevo como un requerimiento procedimental a través del adoptar una nueva decisión, en caso de resultar pertinente, teniendo en cuenta el punto de vista del interesado. Entre otros, Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems»; Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond»; Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations».

[735] Cuestión que el GT29 pone sobre la mesa: *El interesado solo podrá impugnar la decisión o expresar su punto de vista si comprende plenamente cómo se ha tomado y sobre qué base*. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 30.

[736] Bayamlıoğlu, «The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"», 6. En este mismo sentido: *Es decir, el conocimiento e ilustración de las razones que fundamentan la decisión constituye una precondición para impugnar eficazmente la decisión, pues solo de esta manera se asegura la eficacia del derecho a impugnar la decisión. Esto es, no se puede ejercitar eficazmente el derecho a oponerse a la decisión si previamente no se ha ilustrado a los interesados de las razones que motivaron la decisión. Por ende, el derecho a impugnar la decisión presupone el derecho a recibir una explicación, y consecuentemente, la posibilidad de contestación o rechazo de la decisión.* Armada Villaverde y López Bustabad, «El reglamento general de protección de datos ante el fenómeno del "«big data"».

[737] Directiva (UE) 2016/680, de 27 de abril de 2016, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por parte de las autoridades competentes para fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, y a la libre circulación de dichos datos y por la que se deroga la Decisión Marco 2008/977/JAI del Consejo.

decisión: «*van de suyo en el marco de un proceso penal por efecto del derecho de defensa y contradicción y por efecto de la regulación de recursos en la legislación procesal penal, razón por la cual quizá no se vio necesario incluirlo en el dictado literal del texto normativo de la Directiva*»[738], aunque sí están presentes en su considerando 38[739]. Ello también pone de manifiesto que el elemento humano no es un requisito procedimental más[740], conforme al análisis desarrollado anteriormente la intervención humana ha de ser siempre significativa -independientemente de que se exija como componente esencial de la toma de decisiones, *in the loop*, o como medida de salvaguarda bajo requerimiento del interesado, *out of the loop*-.

Si observamos con detenimiento la -ambigua- distribución de este derecho al debido proceso ante las decisiones plenamente automatizadas caben, no obstante, dos formas muy distintas de entender cómo se ejerce el derecho de impugnar la decisión y que parecen haber sido obviadas por las discusiones jurídicas sobre la materia. Tomemos como punto de partida el análisis ya realizado sobre la intervención humana como medida de salvaguarda y su conexión con el derecho de la persona interesada a expresar su punto de vista.

El derecho a obtener intervención humana tiene como fundamento la reevaluación del resultado algorítmico que, de forma legítima -art. 22(2)-, produce un efecto jurídico o significativo en la interesada[741]. Esta reevaluación debe tener en cuenta todos los datos pertinentes, incluyendo cualquier información adicional facilitada por el interesado[742].

---

[738] Guzman Fluja, «Proceso penal y justicia automatizada», 26.

[739] Directiva 680/2016, considerando 38: *El interesado debe tener derecho a no ser objeto de una decisión que evalúe aspectos personales que le conciernen que se base únicamente en un tratamiento automatizado de los datos y que tenga efectos jurídicos adversos que le conciernan o le afecten significativamente. En todo caso, este tipo de tratamiento debe estar sujeto a las garantías apropiadas, lo que incluye informar de forma específica al interesado, así como el derecho a la intervención humana, en particular para que el interesado pueda expresar su punto de vista, obtener una explicación de la decisión adoptada tras dicha evaluación, o ejercer su derecho a impugnar la decisión. Queda prohibida la elaboración de perfiles que dé lugar a la discriminación de personas físicas por razones basadas en datos personales que, por su naturaleza, son especialmente sensibles en relación con los derechos y las libertades fundamentales, con arreglo a las condiciones previstas en los artículos 21 y 52 de la Carta*.

[740] Guzman Fluja, «Proceso penal y justicia automatizada», 27.

[741] Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems», 1.

[742] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 30. Un mecanismo llamativo para impugnar la toma de decisiones basada únicamente en el tratamiento automatizado es el que se ofrece en la *Data Protection Act* de Reino Unido de 2018 para el tratamiento de datos por los servicios

Volviendo sobre los fundamentos de la disposición kafkiana, esta clase de intervención humana garantiza, por un lado, la permeabilidad de la decisión automatizada –a través del derecho a expresar su punto de vista– y, por otro, la responsabilidad sobre la toma de decisiones –a través de una intervención humana significativa por parte del responsable– . En definitiva, la intervención humana como mecanismo de gobernanza en el apartado 3 tiene por objeto solicitar una segunda resolución, una revisión, en la que un agente humano pueda tener en cuenta también el punto de vista del interesado, mejor dicho, debe tenerlo en cuenta siempre que la persona interesada ejerza su derecho a expresar su punto de vista.

Aquí es donde entran en juego las dos posibilidades sobre el derecho a impugnar la decisión:

A) Se trata de un derecho a impugnar la decisión automatizada inicial y, por ende, a solicitar una segunda resolución, una revisión, en la que un agente humano puede -debe- tener en cuenta también el punto de vista del interesado. Impugnar la decisión, obtener intervención humana y expresar su punto de vista forman parte de una misma instancia.
B) O bien, se trata de un derecho a impugnar esa revisión humana previamente solicitada de forma autónoma. Obtener intervención humana y expresar su punto de vista son una instancia distinta a la impugnación de la decisión.

En definitiva, ¿el proceso debido en el artículo 22(3) RGPD establece una única o una doble instancia? A pesar de la ambigüedad de la disposición, de las Directrices del GT29 en este punto y de la ausencia de posiciones doctrinales claras, en esta investigación abogo por la segunda opción (B) sobre la interpretación del Considerando 71 y, es más, sugiero que la omisión en el texto del derecho a una explicación en el artículo 22 RGPD fue un error, no tanto porque no haya lugar a un derecho a una explicación sobre la decisión totalmente automatizada en los derechos de información y acceso regulados por el

de inteligencia. De acuerdo con su artículo 97(4), esta clase de decisiones pueden impugnarse y solicitar o bien una reevaluación de forma análoga al 22(3) RGPD de un humano out of the loop; o bien una nueva decisión de un humano in the loop: (…) *data subject may, before the end of the period of 1 month beginning with receipt of the notification, request the controller— (a)to reconsider the decision, or (b)to take a new decision that is not based solely on automated processing*. Accesible aquí: https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

Reglamento[743], sino porque omitió un derecho a una explicación autónomo sobre una cuestión distinta al tratamiento automatizado: un derecho a una explicación sobre la revisión humana llevada a cabo.

Observemos el orden dado en el 22(3) RGPD: «*como mínimo el derecho a obtener intervención humana por parte del responsable, a expresar su punto de vista y a impugnar la decisión*».

A primera vista, si el derecho a impugnar la decisión consistiese en solicitar una revisión humana, parece lógico pensar que el orden dado hubiere sido el inverso, colocando el derecho de impugnar en primer lugar. Esta cuestión parece aún más evidente en el Considerando 71: «*En cualquier caso, dicho tratamiento debe estar sujeto a las garantías apropiadas, entre las que se deben incluir la información específica al interesado y el 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*. La clave está en la inclusión del derecho a una explicación *después de tal evaluación*»[744], el considerando parece separar la evaluación llevada a cabo con intervención humana y teniendo en cuenta el punto de vista del interesado, la explicación sobre dicha evaluación y el posterior derecho a impugnar la decisión -revisada ya en esa primera instancia- del responsable del tratamiento.

A pesar de que esta medida de salvaguarda -derecho a una explicación después de tal evaluación- no tuvo reflejo en el texto final del artículo 22(3), el orden no fue invertido y, por ende, parece razonable entender que debemos separar la instancia en la que se solicita una reevaluación de la decisión con intervención humana y derecho a expresar su punto de vista y la instancia en la que se impugna la decisión final, no habiendo satisfecho la reevaluación los intereses de la persona interesada.

Esta interpretación gramatical y lógica, a mi modo de ver, es también más garantista con los derechos de las personas interesadas y esa doble instancia no tiene que suponer un esfuerzo desproporcionado para el responsable del tratamiento. Sin embargo, el reconocimiento explícito del derecho a una explicación sobre esa primera instancia

[743] Sobre cuyo alcance ya se ha discutido. Vid. Apartado 2.2. Derecho a la información para las decisiones basadas únicamente en el tratamiento automatizado, ¿derecho a una explicación?, en este mismo capítulo.

[744] También en la versión en inglés: *the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision*.

hubiera resultado más satisfactorio. Un derecho a una explicación sobre la evaluación realizada sería más apropiado para convertir el uso de perfiles para la toma de decisiones automatizada en un espacio público igualitario, plural y abierto en los términos descritos por Turégano[745].

Conforme a lo desarrollado anteriormente, no parece haber dificultades para que, bajo la normativa actual: (1) los derechos de información y acceso deban incluir una "explicación" sobre la lógica realmente empleada en la decisión particular, y no sólo sobre la funcionalidad general del sistema de una toma de decisiones algorítmica y (2) que dicha información para resultar significativa en los términos de los artículos 13(2)(f), 14(2)(g) y 15(1)(h) RGPD, deba resultar útil, inteligible y accionable –respecto del resto de derechos reconocidos– para la interesada[746]. Siguiendo esta línea, Bayamlioglu desarrolla cómo estos derechos pueden contribuir razonablemente al cumplimiento del derecho a impugnar la decisión del artículo 22(3) RGPD[747].

Sin embargo, si a las posibilidades de esta información sobre la decisión adoptada por el algoritmo, sumáramos una explicación sobre por qué las razones expuestas por el interesado conforme a su punto de vista no han sido suficientes para modificar la decisión conforme a una intervención humana de carácter significativo, es indudable que la permeabilidad del proceso decisorio -que su carácter democrático- se vería notablemente incrementada, además de incrementar también la transparencia sobre la intervención humana -*out of the loop*-.

En cualquier caso, el derecho a impugnar la decisión no parece reconocer la resolución de una segunda instancia por una autoridad neutral, como podrían ser las autoridades de control de los EEMM, y de forma implícita parece que su ejercicio debe resolverse por un procedimiento interno del responsable del tratamiento[748]. Y esta es, a mi juicio, una oportunidad perdida para el RGPD[749]. Por un lado, una segunda instancia para impugnar

[745] Turégano Mansilla, «Los valores detrás de la privacidad», 279.

[746] Selbst y Powles, «Meaningful information and the right to explanation», 242.

[747] También desde una perspectiva práctica. Vid. Bayamlıoğlu, «The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"», 8-10.

[748] Geburczyk, «Automated administrative decision-making under the influence of the GDPR – Early reflections and upcoming challenges», 7.

[749] Lo es porque, de hecho, esta posición venía ya manteniéndose por el Consejo de Europa, que abogaba por -primero- un derecho a una reevaluación humana que pudiera apartarse de la decisión algorítmica bajo criterios razonables, y -segundo- por un derecho a impugnar las decisiones algorítmicas ante una autoridad

la reevaluación humana parece perder sentido si es, de nuevo, un procedimiento interno bajo el control del responsable del tratamiento. Por otro lado, ello supondría una reducción de la carga burocrática para el responsable del tratamiento, no sería necesario reconocer una segunda instancia interna y realizar un nuevo procedimiento ante el derecho de impugnar la reevaluación, sino informar apropiadamente sobre las posibilidades de impugnar dicha decisión final conforme a derecho -pongamos ante una autoridad de control o incluso informando sobre vías jurídicas alternativas en función del ámbito de aplicación sectorial de la decisión-.

Aunque el derecho al "debido proceso" -introducido por el conjunto de medidas de salvaguarda mínimas del artículo 22(3)- podría resultar prometedor para permitir impugnar las decisiones algorítmicas basadas únicamente en el tratamiento automatizado y, con ello, favorecer la influencia de las personas interesadas sobre las decisiones que les afectan de forma significativa, hay factores que impiden cumplir con estos objetivos.

La característica ambigüedad de esta disposición es también notable en este apartado, no solo porque estas medidas mínimas no se aplican de forma uniforme a todas las excepciones de los apartados 22(2) y (4)[750], sino porque no ha permitido dilucidar con claridad si el propio 22(3) se trata de una única o doble instancia sobre la que impugnar la toma de decisiones automatizada. El no reconocimiento explícito del derecho a una explicación sobre la reevaluación humana, tal y como se recoge en el considerando 71, también ha limitado el potencial del mismo y, por último, que el derecho de impugnar se limite al ámbito interno del responsable del tratamiento es una oportunidad perdida para poder resolver esta conflictividad en el marco normativo de la protección de datos desde una instancia externa.

4. Reflexiones provisionales sobre el capítulo tercero – tentative thoughts on chapter three

competente. Vid. Consejo de Europa. Consultative Committee of Convention 108 y Council of Europe, «Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data», 5.

[750] Lo cual ha dado lugar a interpretaciones muy distintas de los EEMM, es preciso volver a recomendar el estudio de derecho comparado de Malgieri en este sentido, Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other “suitable safeguards” in the national legislations».

En este apartado se recogen una serie de reflexiones provisionales a modo de cierre de cada capítulo. Aunque algunas de estas reflexiones servirán de apoyo para las conclusiones de esta investigación, el objetivo de este apartado no es exponer dichas conclusiones propiamente, sino resaltar de forma telegráfica algunos aspectos clave resultado del análisis realizado en cada capítulo.

- La Comisión Europea reveló en los trabajos preparatorios de la Directiva de 1995 dos temores fundamentales sobre la toma de decisiones automatizada y la elaboración de perfiles: (1) un temor a la pérdida del control humano de los responsables del tratamiento sobre las decisiones que adoptan; (2) y un temor a la pérdida de control humano de las personas interesadas sobre las decisiones que les afectan directamente.
- La interpretación propuesta en esta investigación trata de poner en valor tres aspectos normativos fundamentales para enfrentar esos dos temores: la responsabilidad sobre el tratamiento [responsabilidad], la capacidad de la persona interesada para influir sobre el tratamiento [permeabilidad] y la posibilidad de observar por parte de terceros que dicho tratamiento es permeable y responsable [transparencia].
- El artículo 22 contiene dos mecanismos de gobernanza basados en la intervención humana claramente diferenciados: (1) intervención humana como componente esencial de la toma de decisiones -22(1)- que el responsable del tratamiento debe introducir para evitar la prohibición general de forma previa a la producción de efectos para el interesado; (2) intervención humana como medida de salvaguarda -22(3)- posterior a la toma de decisiones basada únicamente en el tratamiento automatizado bajo requerimiento del interesado.
- La intervención humana en el RGPD debe tener un carácter significativo. Conforme a las Directrices del GT29 -refrendadas por el CEPD- esta clase de intervención exige [a] llevarse a cabo por persona autorizada y competente para modificar la decisión, [b] realizarse sobre un análisis que tenga en cuenta todos los datos disponibles y [c] no conllevar aplicación rutinaria de los resultados algorítmicos. En el caso de la intervención humana como medida de salvaguarda debe, además, [d] tener en cuenta la información facilitada por el interesado.

- Sobre los derechos de información y acceso de la elaboración de perfiles, pueden distinguirse tres dimensiones (1) la información sobre la propia actividad de perfilado entre los fines del tratamiento; (2) el perfil generado, esto es, el dato personal en sí y; (3) la lógica y consecuencias de dicha actividad. El RGPD no aporta la suficiente seguridad jurídica a la hora de interpretar en qué clase de elaboración de perfiles debe informarse sobre la lógica aplicada y sus consecuencias -al margen de cuando constituyan decisiones basadas únicamente en el tratamiento automatizado-.
- El derecho a una explicación -tal y como fue concebido en el considerando 71- no fue incluido en la parte dispositiva del RGPD, por ello, lo verdaderamente relevante es poder determinar con criterios interpretativos claros el alcance de los derechos de información y acceso sobre la lógica algorítmica aplicada y sus consecuencias -13(2)(f), 14(2)(g) y 15(1)(h)-. Parece razonable entender que estos derechos obligan también a revelar información sobre la lógica realmente empleada en la decisión particular, y no sólo sobre la funcionalidad general del sistema de una toma de decisiones algorítmica.
- El RGPD impone que el responsable del tratamiento haga un uso responsable, transparente, leal y exacto, entre otros, de los datos personales con independencia de los métodos tecnológicos que utilice, por tanto, no cabe ampararse en formas de opacidad inherentes a los modelos algorítmicos para eludir derechos de información y acceso de los interesados u obligaciones de demostración del cumplimiento normativo.
- La posibilidad de rebatir, refutar o discutir las inferencias algorítmicas y las decisiones automatizadas por las personas interesadas afectadas por las mismas define el nivel de permeabilidad del RGPD. En este sentido, el derecho de rectificación -16- aplicable a la elaboración de perfiles y la salvaguarda de impugnación de decisiones automatizadas -22(3)- pueden considerarse derechos “hermanos”.
- El derecho a impugnar la decisión basada únicamente en el tratamiento automatizado puede considerarse la columna vertebral de las medidas de salvaguarda para los derechos y libertades de las personas interesadas establecidas por el RGPD para esta clase de decisiones. En este sentido, el ejercicio de medidas de salvaguarda -22(3)- y los derechos de información y acceso tienen un carácter

instrumental respecto del derecho a impugnar decisiones basadas únicamente en el tratamiento automatizado.

As a method of recapping each chapter, this section presents a number of tentative thoughts. The research's conclusions will be supported by some of these insights, but the purpose of this section is not to present those conclusions in their entirety. Rather, it aims to emphasize certain key aspects that came out of the analysis done in each chapter in a telegraphic manner.

- In the preparatory works for the 1995 Directive on data protection, the European Commission revealed two fundamental fears about automated decision-making and profiling: (1) a fear concerning the loss of human control by data controllers over the decisions they adopt; (2) and a fear concerning the loss of human control by data subjects over decisions that directly affect them.
- The interpretation put forth in this study aims to draw attention to three crucial normative aspects that can help address these two concerns revealed by the European Commission: accountability over processing [accountability], the data subject's ability to influence processing [permeability], and the possibility for third parties to observe that processing is permeable and accountable [transparency].
- Article 22 contains two clearly differentiated governance mechanisms based on human intervention: (1) human intervention as an essential component of decision-making -22(1)-, which the controller must introduce to prevent the general prohibition prior to the production of effects for the data subjects; (2) human intervention as a safeguard measure -22(3)-, which comes after decision-making based solely on automated processing at the request of the data subject.
- In the GDPR, human intervention must be meaningful. According to the WG29 Guidelines -endorsed by the EDPB- this kind of intervention requires [a] to be carried out by a person authorised and competent to change the decision, [b] to be based on an analysis that takes into account all available data and [c] not to involve routine application of algorithmic results. For human intervention as a safeguard measure it must, in addition, [d] take into account the information provided by the data subject.

- Regarding the information and access rights for profiling, three dimensions can be distinguished (1) the information about the profiling activity itself among the purposes of data processing; (2) the profile generated, i.e. the personal data itself and; (3) the logic and consequences of the profiling activity. Except when profiling constitutes a decision based solely on automated processing, the GDPR does not provide adequate legal certainty when interpreting what form of profiling the logic utilized and its effects must be informed.
- The right to an explanation -as conceived in Recital 71- was not included in the provisions of the GDPR. Therefore, what is relevant is to be able to determine with clear interpretative criteria the scope of the rights of information and access on the algorithmic logic applied and its consequences - 13(2)(f), 14(2)(g) and 15(1)(h) -. It seems reasonable to understand these rights as also requiring disclosure of information about the logic actually employed in a particular decision, and not only about the overall functionality of an algorithmic decision-making system.
- No matter the technological tools employed, the controller is required by the GDPR to use personal data, among other things, in a responsible, transparent, fair, and accurate manner. Therefore, inherent forms of opacity in algorithmic models cannot be used to circumvent data subjects' rights of information and access or obligations to demonstrate compliance.
- The degree to which data subjects may challenge, contest or argue the algorithmic inferences and automated decisions that affect them determines the level of permeability of the GDPR. In this sense, the right to rectification -16- applicable to profiling and the safeguard to contest automated decisions -22(3)- might be seen analogue rights.
- The right to contest the decision based solely on automated processing can be considered as the backbone of the safeguards for the rights and freedoms of data subjects established by the GDPR for this kind of decisions. In this regard, the exercise of other safeguards -22(3)- and the rights of information and access are instrumental to exercise of the right to contest decisions based solely on automated processing.

CAPÍTULO 4. LA INTERVENCIÓN HUMANA Y EL PRINCIPIO DE RESPONSABILIDAD EN EL TRATAMIENTO DE DATOS PERSONALES: UN ENFOQUE BASADO EN LA EVIDENCIA A TRAVÉS DE LA EVALUACIÓN DE IMPACTO. UNA PROPUESTA DESDE LA MEDICINA PREVENTIVA[751]

[751] Quiero agradecer a Paul de Hert que en mi visita a su ciudad natal (Antwerp-Amberes) y tras la lectura del primer borrador del trabajo que deseaba desarrollar durante mi estancia en la VUB, identificó rápidamente el vínculo entre la intervención humana y la responsabilidad que yo había esbozado, me obligó a explicarlo y sistematizarlo mucho mejor y me ayudó a explotarlo aportando argumentos que han visto la luz en algunos de los trabajos que firmamos juntos.

CAPÍTULO 4. LA INTERVENCIÓN HUMANA Y EL PRINCIPIO DE RESPONSABILIDAD EN EL TRATAMIENTO DE DATOS PERSONALES: UN ENFOQUE BASADO EN LA EVIDENCIA A TRAVÉS DE LA EVALUACIÓN DE IMPACTO. UNA PROPUESTA DESDE LA MEDICINA PREVENTIVA

Al comienzo del anterior capítulo se advertía de que los tres pilares desarrollados se construyen sobre la base de los derechos de las personas interesadas recogidos en el ecosistema normativo del RGPD. Aunque esta perspectiva no deja de ser acertada -el artículo 22 es parte de los derechos de los interesados, al igual que los derechos de información y acceso o el derecho de rectificación-, la ubicación de estos derechos lleva habitualmente a pensar que el RGPD es, esencialmente, una norma sobre la que reconocer y ejercer derechos individuales.

No solo el debate sobre la transparencia y explicabilidad algorítmica ha desviado el interés sobre mecanismos de gobernanza igualmente útiles y presentes en la regulación de la toma de decisiones automatizada, sino que también el énfasis sobre los derechos individuales correspondientes a las personas interesadas, habitualmente ya de por sí en una posición vulnerable[752], ha desviado la atención sobre una la piedra angular del RGPD: la responsabilidad sobre el cumplimiento normativo y su demostración.

Este capítulo final parte de una preocupación muy extendida en la literatura jurídica y científica, ¿pueden los mecanismos de gobernanza basados en la intervención humana garantizar los fines normativos a los que responde la inclusión de la supervisión humana? Y, en particular, ¿es posible una intervención humana significativa en los términos exigidos por el RGPD? Las respuestas que trato de aportar a este debate no tienen, por supuesto, la intención de dar por cerrado el mismo. Pero sí apuntan en una dirección: es necesario un enfoque basado en la evidencia para adoptar y diseñar mecanismos de intervención humana que resulten útiles a los fines normativos para los que se incluye esta intervención.

A partir de esta reflexión, se indaga en el papel central otorgado al principio de responsabilidad en el modelo normativo -o ecosistema- del RGPD y en la que es, probablemente, la obligación jurídica más relevante en el tratamiento de alto riesgo de datos personales: la evaluación de impacto de protección de datos (EIPD, en adelante).

[752] Cobbe y Singh, «Reviewable Automated Decision-Making», 2.

En este análisis se pone de manifiesto cómo la clase de justificaciones exigidas por la EIPD para demostrar el cumplimiento con el RGPD en la regulación de la toma de decisiones automatizada contienen un enfoque basado en la evidencia que puede resultar de extraordinaria utilidad. Por último, se analiza cómo la intervención humana se relaciona en el ecosistema del RGPD con la EIPD, demostrando que -aun con severas limitaciones- podemos encontrar en el RGPD la obligación de diseñar y aplicar una intervención humana significativa y demostrable para los procesos de toma de decisiones automatizada basada en la elaboración de perfiles.

1. ¿Es posible una intervención humana significativa? Partiendo de las críticas a la intervención humana para reivindicar una supervisión humana basada en la evidencia

En este apartado es necesario abordar las críticas que el derecho a la intervención humana ha despertado como remedio normativo para la toma de decisiones automatizada. Y es que, habitualmente, desde el ámbito legislativo se ha asumido que incluir humanos en la toma de decisiones, especialmente como autoridad final, solucionaría los problemas asociados al uso de sistemas automatizados[753]. Sin embargo, ya se ha señalado anteriormente que, entre los sistemas automatizados, sean totalmente automatizados o sirvan como sistemas de apoyo a la toma de decisiones, los problemas son similares[754].

Ahora bien, aquí, estas limitaciones deben ser abordadas desde una perspectiva jurídica, por lo que debemos preguntarnos cuál es la justificación para introducir un derecho a la intervención humana y cómo estas limitaciones pueden afectar a la misma, determinando si ésta puede o no tener un carácter significativo.

Para ello, acudiremos a los fundamentos esgrimidos en los anteriores apartados, que pueden resumirse en: uno, la introducción de la intervención humana como componente esencial de la toma de decisiones con el objetivo de garantizar un tratamiento automatizado responsable, lícito, leal y exacto en la toma de decisiones; dos, intervención humana como componente esencial de la toma de decisiones como fundamento de la

[753] Wagner, «Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems», 107.

[754] Citron, «Technological Due Process», 1267.

dignidad humana[755]; y tres, la intervención humana como medida de salvaguarda para facilitar la participación de las personas interesadas en las decisiones basadas únicamente en el tratamiento automatizado que les afectan de forma significativa.

Comenzando por este último fundamento, Huq analiza el mismo y encuentra motivos para ser escéptico sobre el derecho de intervención humana. Así, se refiere a la posibilidad de que un individuo afectado por la decisión aporte información y razones a un ser humano por las cuales debería tomarse una solución favorable a su caso[756]. Se asume que una revisión *ex post* ayudaría a resolver falsos negativos de personas perjudicadas por el sistema algorítmico, sin embargo, no hay ninguna razón empírica para esperar que la incorporación de una revisión humana aumente la precisión general, al contrario, la revisión humana parece aumentar generalmente las tasas de error netas[757]. En otras ocasiones, cuando los parámetros relevantes estén tasados por una norma, Huq cuestiona qué valor podría aportar una revisión humana a un proceso decisorio que debe aceptar solo razones tasadas, en el que la capacidad de influencia debería ser nula[758]. En este sentido, se ha planteado la posibilidad de que en el futuro sea posible utilizar un proceso algorítmico para demostrar la legitimidad de una decisión tomada por un ser humano o por una máquina más efectiva que cualquier revisión humana de la decisión en cuestión[759]. Para Huq, igualmente, en lugar de dedicar recursos a una revisión humana individualizada, habría de dedicarlos a una revisión sistemática algorítmica[760].

[755] Sobre el alcance de este fundamento, en el capítulo anterior he argumentado que la restauración de la dignidad humana a través de la intervención humana se traduce en el RGPD en que la aplicación rutinaria de los resultados algorítmicos generados automáticamente sin influencia real por el agente humano que las supervisa no puede considerarse intervención significativa, en línea con Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 23.

[756] Huq, «A Right to a Human Decision», 661.

[757] Huq, 666-67.

[758] Huq, 669.

[759] Kuner et al., «Machine learning with personal data: is data protection law smart enough to meet the challenge?», 2.

[760] Huq, «A Right to a Human Decision», 671; también en este sentido, Cotino, «Ética en el diseño y confiable para el desarrollo de la robótica, inteligencia artificial y el big data y su utilidad desde el derecho», 43.

No obstante, al margen de esta visión tan crítica, hay también parte de la doctrina que reclama el elemento humano por valores que asociamos a las relaciones humanas en sí, como la capacidad de escucha o la empatía[761].

Acerca de la restauración de la dignidad como fundamento de la intervención humana, el sesgo de automatización hace difícil que podamos confiar sistemáticamente en los agentes humanos para solucionar o mitigar los problemas asociados a los sistemas automatizados[762]. Las personas pueden creer que un sistema automatizado está mejor capacitado que ellas, y con el tiempo pueden perder sus habilidades para evaluar los resultados de los algoritmos[763]. Es más, con el desarrollo de algoritmos cada vez más complejos y no interpretables por humanos, la capacidad para evaluar sus decisiones es todavía más cuestionable[764].

Además, no solo factores puramente tecnológicos afectan a la capacidad para intervenir en la toma de decisiones automatizada, el régimen de responsabilidad jurídica aplicable o la forma de organización empresarial o institucional pueden influir en dicha capacidad[765]. Todo ello, cuestiona la viabilidad de una intervención humana que pueda garantizar que los humanos mantengan el papel principal en la constitución de otros seres

[761] Tamò-Larrieux, «Decision-making by machines: Is the 'Law of Everything' enough?», 5. Pongamos que, efectivamente, la intervención humana en la fase de revisión no se traduce en una mejora de la precisión general del sistema automatizado y así es constatado por el responsable del tratamiento. Sin embargo, éste constata a través de encuestas de satisfacción que la intervención humana da lugar a una satisfacción mayor en el proceso por parte del interesado, o que da lugar a una reducción de la litigiosidad. ¿Podríamos considerar tal intervención como significativa a pesar de no mejorar la precisión del sistema?

[762] Yeung, «Algorithmic regulation: A critical interrogation», 516.

[763] Citron, «Technological Due Process», 1272. A pesar de ello, no obstante, los seres humanos en el proceso de toma de decisiones no deben ser vistos como meros objetos pasivos, la agencia humana también desempeñará un papel clave sobre cuándo y cómo se utilizarán los sistemas de IA en diferentes contextos sociotécnicos. Las conclusiones de Sandhu y Fussey cuestionan la idea de que la experiencia y las habilidades de los agentes de policía vayan a ser sustituidas rápidamente por las tecnologías predictivas; el trabajo policial tiende más bien a cambiar lentamente a pesar de la adopción de la tecnología, lo que refleja los ritmos preexistentes de las estructuras policiales. Vid. Sandhu y Fussey, «The 'uberization of policing'? How police negotiate and operationalise predictive policing technology», 78-79.

[764] Noto La Diega, «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information», 19.

[765] Para Brkan la disposición de un humano a reevaluar y modificar la decisión dependerá en gran medida de su responsabilidad por la decisión final, vid. Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond».

humanos y sea, por tanto, una intervención significativa en este sentido que no vulnere la dignidad humana[766].

En cuanto al primer fundamento -la introducción de la intervención humana como componente esencial de la toma de decisiones con el objetivo de garantizar un tratamiento automatizado responsable, lícito, leal y exacto en la toma de decisiones-, parte de la doctrina considera que nada garantiza que la intervención humana mejore la decisión algorítmica tomada sobre el interesado, es más, podría darse el caso de que la persona empeore la decisión alternativa que habría tomado el sistema de manera puramente automatizada[767]. Para Huq, el hecho de que un sistema tome decisiones defectuosas no implica que una decisión humana sea mejor[768].

Parte de la doctrina considera que el juicio humano debe mantenerse debido a una supuesta capacidad intuitiva, propiamente humana, que permitiría identificar problemas y errores de las máquinas[769], sin embargo, las proclamas por la intuición humana no parecen estar -siempre- suficientemente justificadas[770].

La evidencia también señala que las decisiones humanas no solo están sesgadas, de forma más o menos similar a como lo están las máquinas, sino que también se ven afectadas por

---

[766] A este respecto, es interesante la reflexión de Cabitza que invita a abandonar el objetivo inalcanzable de una "IA centrada en el ser humano", por el objetivo de minimizar los sesgos individuales inducidos por la IA, como el sesgo de automatización, que pretende lograr una "humanidad descentrada de la IA", en Cabitza, «Many say that AI can outperform human doctors. Is it true?».

[767] Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems», 5; Roig, *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*, 57.

[768] Huq, «A Right to a Human Decision», 671. Lo cual podría llevarnos al debate de si es aceptable utilizar tecnologías que tomen decisiones defectuosas por el hecho de que los humanos que han tomado dichas decisiones habitualmente lo han hecho de forma igualmente fallida.

[769] Brennan-Marquez y Henderson, «Artificial Intelligence and Role-Reversible Judgment», 146. También en de Montalvo, «¿Puede la máquina sustituir al hombre? Una reflexión jurídica sobre el ojo clínico y la responsabilidad en tiempos de Big Data».

[770] McGuire muestra cómo la necesidad de la intervención humana en la labor policial se reivindica a menudo sobre la base de "intuiciones" y "normas morales" muy cuestionables, en McGuire, «The laughing policebot: automation and the end of policing», 29-30. Más acertado parece el enfoque de Binns, que se pregunta si la discrecionalidad es verdaderamente humana. Así, argumenta que los sistemas automatizados no pueden proporcionar una justicia individual caso por caso, en el sentido de que un nuevo caso podría tener que ser tratado de forma diferente a uno anterior aparentemente idéntico. Y el juicio humano lleva a incoherencias al no tratar por igual dos casos que son iguales en aspectos relevantes (por aparición del ruido, como veremos a continuación). En sus propias palabras: *While algorithmic systems arguably serve one dimension of justice very well (consistency), they fail to respect individual justice and are no guarantee of nondiscrimination. Similarly, even though human judgment is necessary for the individual dimension of justice, it also risks conflicting with both consistency and nondiscrimination*. Vid. Binns, «Human Judgment in algorithmic loops: Individual justice and automated decision-making», 6-7.

el fenómeno definido como ruido. Kahneman et al. definen el ruido como la variabilidad no deseada en el juicio humano, y muestran cómo conduce a errores significativos que no deben ser ignorados, al menos si se pretende solucionar los mismos[771].

Siendo este aspecto difícilmente rebatible, no puede obviarse, por un lado, que la definición de lo "no deseable" nunca es neutra y, por otro, que el ruido no puede llevarnos a proponer soluciones simplistas de escasa evidencia[772]. Tampoco parece razonable asumir de forma acrítica la premisa de que, estando las decisiones humanas igualmente sesgadas a las decisiones automatizadas, no resultará satisfactoria para la minimización de dichos sesgos la introducción de intervención humana en el proceso decisorio. El impacto del sesgo algorítmico es muy diferente del de un decisor humano, Gandy señala que el hecho de que el número de decisiones tomadas por estos sistemas pueda ser varios órdenes de magnitud mayor que las tomadas por los decisores humanos, sugiere que los daños acumulativos, así como los catastróficos, surgirían simplemente en función de la escala[773].

Son varios los puntos sobre los que debemos centrar nuestra atención aquí. En primer lugar, la definición de lo que es mejor o peor no es, otra vez, neutra. Lo que es mejor en términos de exactitud podría calcularse sobre una base estadística que resulta discriminatoria. Es más, dos errores, pongamos un falso positivo y un falso negativo, pueden tener consecuencias muy diferentes en función del contexto de aplicación del sistema. Además, como en todo contexto normativo complejo, a la hora de entender cómo introducir la intervención humana, encontraremos distintos valores compitiendo entre sí[774]. En este caso, es evidente que los distintos principios del RGPD compiten entre sí en muchos ámbitos -por ejemplo, el principio de minimización de datos con el de exactitud a la hora de entrenar un algoritmo-, y también en la intervención humana existirá esa tensión que ya hemos podido ejemplificar entre la licitud, lealtad y exactitud. Lo

[771] Vid. Kahneman, Sibony, y Sunstein, *Noise: A Flaw in Human Judgment*.

[772] Entre las cuales, podemos encontrar las que el propio Kahneman propuso anteriormente: *Uncomfortable as people may be with the idea, studies have shown that while humans can provide useful input to formulas, algorithms do better in the role of final decision maker. If the avoidance of errors is the only criterion, managers should be strongly advised to overrule the algorithm only in exceptional circumstances*. Vid. Kahneman et al., «Noise: How to overcome the high, hidden cost of inconsistent decision making».

[773] Gandy, «Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems», 39.

[774] Brennan-Marquez y Henderson, «Artificial Intelligence and Role-Reversible Judgment», 142.

relevante, desde mi punto de vista, es poder identificar esta tensión y justificar unos u otros "sacrificios".

En segundo lugar, debe superarse una visión simplista muy habitual en este debate que aboga por una ficticia lucha entre el juicio humano y el juicio de la máquina[775]. No solo por el hecho de que muchas de estas comparaciones se realizan sobre bases empíricas más que cuestionables con poca o ninguna utilidad práctica[776], sino también porque no parece que el horizonte a corto y medio plazo vaya a ser la eliminación de los humanos de la toma de decisiones en aspectos que nos conciernen de forma directa[777]. Tampoco parece que las máquinas estén preparadas para asumir esta responsabilidad[778].

Otra cuestión que no puede perderse de vista es que todos los defectos presentes en el juicio y toma de decisiones humanos, no se dan de forma exclusiva en la fase de uso e implementación de un modelo. Es más, la doctrina más crítica reconoce que el diseño y desarrollo de los modelos está plagado de intervención humana[779]. Salvo que se considere que, por alguna razón, el juicio de quienes codifican el algoritmo es superior al de los y las profesionales que puedan intervenir en la fase de uso del modelo -pongamos médicas, jueces o personal de recursos humanos-, no parece razonable excluir la intervención de unos u otros en el ciclo de vida del modelo en atención a los defectos del juicio humano. Habrá que atender, por supuesto, a las limitaciones que se presentan en cada fase y establecer, en consecuencia, mecanismos jurídicos adecuados para que la supervisión humana responda a los objetivos normativos fijados.

Hildebrandt propone un cambio de perspectiva retórico para la supervisión humana, desde el '*human in the loop*' al '*machine in the loop*', de forma que lo humano represente el

---

[775] En estos términos se plantea el debate, por ejemplo, Huq, «A Right to a Human Decision»; también en Tamò-Larrieux, «Decision-making by machines: Is the 'Law of Everything' enough?»

[776] Cabitza, «Many say that AI can outperform human doctors. Is it true?»

[777] Como hemos visto anteriormente, la supervisión humana parece ser un aspecto innegociable en toda iniciativa de regulación de los sistemas de IA considerados de alto riesgo.

[778] Entiéndase la ironía.

[779] En palabras de Huq: *Recall that the design, testing, and implementation of machine-learning tools are all thoroughly imbricated with purposeful human choice and intentionality. Human intentions necessarily guide the choice between supervised and unsupervised models; the process of feature selection; the selection of training data; and the ongoing process of refinement and calibration toward an optimal classifier. Much of this intentional human action is necessarily oriented by an understanding of the ends that the machine will serve. The dearth of reasoned judgments in machine decisions, therefore, is something of an optical illusion*. Huq, «A Right to a Human Decision», 675.

aspecto central en la toma de decisiones y sea la máquina la que integremos en dicho *loop*[780]. En términos similares, Cabitza et al. (a partir del estudio de las leyes de Garry Kasparov sobre equipos colaborativos humano-máquina) proponen centrarse en cómo los agentes humanos y la IA son un mismo sistema sociotécnico, integrado en los contextos sociales y profesionales preexistentes, y superar una valoración basada en la evaluación de su rendimiento como agentes únicos y aislados[781]. ¿Cómo podría integrarse este cambio de perspectiva en la forma de entender la intervención humana significativa?

Si queremos concebir al agente humano y la IA como parte de un mismo sistema sociotécnico, por un lado, la intervención humana no puede construirse sobre la desconfianza o la fe ciega en la tecnología, sino sobre un conocimiento empírico de los estándares de calidad, limitaciones e impacto de la tecnología. El diseño y desarrollo defectuoso de un sistema, incluyendo el uso de metodologías cuestionables[782], no será resuelto por ninguna clase de intervención humana[783] y, por ende, ésta tampoco contribuiría a la responsabilidad en el cumplimiento del RGPD. Lo cual nos lleva también a la siguiente consideración: la intervención humana por sí misma no puede ser suficiente para conseguir una adecuada supervisión sobre los sistemas automatizados[784].

---

[780] Vid. Hildebrandt, «Comments on White Paper on AI (EC)».

[781] Cabitza, Campagner, y Sconfienza, «Studying human-AI collaboration protocols: the case of the Kasparov's law in radiological double reading», 18.

[782] Aquí puede incluirse, por ejemplo, la falacia de la frecuencia base a la que hizo referencia Korff durante los trabajos preparatorios del RGPD. Esta falacia se fundamenta en el hecho matemáticamente inevitable de que, si se buscan casos muy raros en un conjunto de datos muy grande, por muy bien que se diseñe el algoritmo, siempre se acabará teniendo un número excesivo de falsos positivos o de falsos negativos. Lo ilustra con el siguiente ejemplo: *Statisticians know this. Epidemiologists know this: they know that it is effective to screen all women over the age of 50 for breast cancer, because in that group there is a sufficiently high incidence of that affliction. But it is not effective to screen all women over the age of, say, 15, because that would thrown up enormous numbers of "false positives", which would deplete hospital resources. Exactly the same applies in antiterrorist screening based on profiles: there are (thank God) simply not enough terrorists in the general population, or even in smaller populations (say, all Muslims in the UK of Pakistani or Saudi origin), to make the exercise worthwhile. The police and the security services would be chasing thousands of entirely false leads, while some real terrorists would still slip through the net*. Vid. Korff, «Comments on Selected Topics in the Draft EU Data Protection Regulation».

[783] Es más, lo probable es que la intervención empeore el rendimiento del sistema por sí defectuoso, así lo expresa Selbst en el ámbito de la prevención criminal: *In these proposals, predictive algorithms become similar to any other direct observations of suspicious behavior, after which the police use their discretion to decide whether to act. But such a process only serves to double the sources of disparate impact. Not only will the effects of the algorithms' disparate impact go unrecognized by police and be treated as a neutral fact, but the discretion that the "neutral algorithm" is supposed to solve again becomes a part of the overall decision*. Selbst, «Disparate Impact in Big Data Policing», 160.

[784] Esta idea aparece implícita en la argumentación de la Sentencia de la *England and Wales Court of Appeal*, de 11 de agosto de 2020, en el caso *R (on the application of Edward Bridges) v the Chief Constable of South Wales Police*. [2020] EWCA Civ 1058, caso núm. C1/2019/2670. Este caso se refiere a la licitud

Por otro lado, tampoco puede celebrarse la intervención humana que produce resultados discriminatorios o inexactos y, por ende, no puede edificarse la intervención humana significativa sobre argumentos pretenciosos o casi mágicos sobre el juicio humano[785]. Al contrario, la intervención humana debe edificarse sobre evaluaciones empíricas en las que se destaquen los aspectos en los que la supervisión humana contribuye al cumplimiento del ordenamiento jurídico en el contexto social e institucional particular en el que se introduce el sistema -*machine in the loop*-. De esta forma, la intervención humana debe configurarse como una pieza más en la cadena organizativa para conseguir un tratamiento automatizado de los datos personales lícito, leal, exacto y responsable, entre otros, exigido por el RGPD.

Dados los argumentos expuestos que, sin duda, merecerían un análisis y debate jurídico en mayor profundidad, en esta investigación propongo adoptar un *enfoque basado en la evidencia*.

Las críticas aquí recogidas ponen de manifiesto muchas de las limitaciones que la intervención humana padece como remedio normativo, no obstante, dichas críticas también pueden resultar endebles en ciertos aspectos, muy especialmente cuando ponemos en consideración las alternativas disponibles, es decir, la ausencia de intervención humana para la implementación de esta clase de sistemas. En este contexto, en el que podríamos estar obligados a elegir entre una solución imperfecta y otra igualmente imperfecta, la solución adoptada debe fundamentarse en la evidencia de que se trata de la más satisfactoria. Sin perder de vista que siempre existirá alternativa cuando esta perspectiva basada en la evidencia arroje, si se da el caso, soluciones igualmente insatisfactorias: la posibilidad de escoger no hacer uso del sistema automatizado. Y, por

---

del uso de la tecnología de reconocimiento facial automatizado en vivo, que utiliza un software llamado AFR Locate (AFR), por parte del Cuerpo de Policía de Gales del Sur. Al contrario que el tribunal de primera instancia, el de apelación consideró que un "mecanismo de seguridad humano", el hecho de que una persona compruebe el resultado antes de actuar sobre él, no era suficiente para cumplir con los deberes de procedimiento establecidos por la Public Sector Equality Duty (PSED) en la Ley de Igualdad de 2010, que exigen la adopción de medidas razonables para hacer averiguaciones por parte de una autoridad pública acerca del impacto potencial de una decisión o política propuesta sobre las personas a las que afectará, teniendo en cuenta características relevantes, en particular, la posible discriminación racial y sexual -par. 181-. El Tribunal tuvo en cuenta el hecho de que los seres humanos también cometen errores -par. 184-185-. Pero no sólo, en este caso el agente humano no pudo comprobar las personas que fueron captadas por la tecnología AFR pero cuyos datos fueron luego borrados casi inmediatamente, tampoco si los datos de entrenamiento estaban sesgados -par. 191 y 193-.

[785] No parece una buena idea establecer estándares normativos para definir la intervención humana significativa que estén por encima de las capacidades humanas.

último, sin perder de vista que la intervención humana por sí misma no puede ser suficiente para conseguir una adecuada supervisión sobre los sistemas automatizados.

¿Podemos considerar que estas conclusiones estén, de algún modo, reflejadas en el ecosistema normativo del RGPD? En el próximo apartado, a partir del análisis del principio de responsabilidad como principio nuclear de la normativa de protección de datos, trato de aportar argumentos para tomar por buena esta hipótesis, sin obviar el margen de mejora de la normativa vigente.

2. Enfoque basado en la evidencia en el RGPD. Principio de responsabilidad -*accountability*-: desplazando la carga desde la persona interesada hacia el responsable del tratamiento

Al indagar en los orígenes sobre la relación entre la privacidad y la toma de decisiones automatizada, he tenido ocasión de hacer referencia a palabras de Turégano en las que expresaba que, ya con la Directiva 95/46/CE de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respeta al tratamiento de datos personales y a la libre circulación de estos datos (en adelante DPD), el foco sobre el control del interesado acerca de la información a él que se revela o deja de revelar, se desplaza hacia un control colectivo sobre la justificación de los procesos que recopilan, procesan y usan los datos[786]. Y el RGPD no hizo sino reforzar este control colectivo basándose en la responsabilidad del cumplimiento normativo[787].

Uno de los cambios normativos más relevantes entre la Directiva de 1995 y el Reglamento de 2016 es la introducción de la "*accountability*" como principio nuclear de la normativa de protección de datos[788]. Este concepto se define por Bovens como: «*una relación socio-normativa entre un actor y un foro, en la que el actor tiene la obligación de explicar y justificar su conducta, el foro puede plantear preguntas y emitir juicios, y el actor puede sufrir consecuencias a partir de ello*»[789]. Puede decirse que este concepto es una idea

[786] Turégano Mansilla, «Los valores detrás de la privacidad», 276.

[787] Aunque no en todos los aspectos, Mayer-Schönberger destaca que el enfásis sobre el consentimiento como base legitimadora del tratamiento en el RGPD es un paso atrás en este sentido. Vid. Mayer-Schönberger, «Paradigm shift».

[788] Traducido en la versión española como "principio de responsabilidad proactiva", en este texto me referiré al principio de responsabilidad.

[789] Bovens, «Analysing and Assessing Accountability: A Conceptual Framework», 452.

fundamental en nuestras democracias; en la democracia representativa, los electores delegan en los representantes el ejercicio del poder de decisión institucional y, a cambio, los representantes deben rendir cuentas ante los electores de diferentes formas[790]. De forma análoga, en la normativa de protección de datos los responsables del tratamiento utilizan la información relativa a los interesados, siempre que cumplan y demuestren el cumplimiento de las normas del RGPD. El ejercicio de esta facultad está sujeto, a su vez, a la rendición de cuentas ante las propias personas interesadas o incluso ante las autoridades públicas competentes.

Sobre el concepto de *accountability*, la clave no está en la imposición de sanciones, sino en el proceso de rendición de cuentas[791]. Para que este proceso resulte efectivo, la información proporcionada por el actor debe apoyar la deliberación y el debate efectivos por parte del foro y la imposición de consecuencias previstas por parte del foro al actor –como recursos legales o intervenciones para corregir el mal funcionamiento del proceso, pensemos en el derecho de rectificación, incluyendo también la posibilidad del establecimiento de sanciones, por supuesto, que corresponde a las autoridades-[792]. La responsabilidad puede definirse, por tanto, como un proceso activo de generación de conocimiento destinado a ser escrutado para restablecer la relación actor-foro haciendo más factible la evaluación de la conducta de aquél[793]. Su principal valor radica en la capacidad de modificar el comportamiento del actor, ajustándolo a las normas definidas para la relación actor-foro, porque el actor sabe que debe justificarlo y las consecuencias de su incumplimiento. Es irrelevante que los efectos del incumplimiento se evalúen a posteriori, dado que la rendición de cuentas tiene un efecto ex ante sobre el comportamiento del actor desde el momento en que éste sabe que está obligado a explicarlo y justificarlo.

[790] Incluyendo aquí formas directas de control por parte de los electores, como la celebración periódica de elecciones, o formas indirectas como la fiscalización de las decisiones institucionales por el poder judicial a través de los mecanismos legalmente establecidos.

[791] Más ampliamente sobre la responsabilidad como mecanismo de gobernanza y su encaje en la normativa de protección de datos, vid. Hert y Lazcoz, «When GDPR-principles blind each other. Accountability, not transparency, at the heart of algorithmic governance».

[792] Vid. Cobbe, Lee, y Singh, «Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems».

[793] Esta definición es de Paul de Hert para el estudio arriba citado.

En la regulación del perfilado y la toma de decisiones automatizada, esta cuestión es de vital importancia puesto que estas actividades pusieron de manifiesto desde el inicio un problema de desequilibrio de poder entre el individuo y las grandes corporaciones públicas y privadas -volvemos sobre el *database privacy problem* de Solove-, acrecentado por el desarrollo tecnológico de las últimas décadas[794]. Es de vital importancia porque la responsabilidad sobre el cumplimiento normativo y su demostración nos habla fundamentalmente de poder, en concreto de empoderar a aquellos que, de otro modo, estarían desprovistos del mismo -frente a la toma de decisiones basada en el perfilado y tratamiento de sus datos personales-, exigiendo que los que ejercen el poder sobre los individuos atomizados rindan cuentas de su conducta[795].

A continuación, se indagará en la forma concreta que ha adoptado la introducción de este principio en el ecosistema del RGPD y, en particular, en la regulación de la toma de decisiones automatizada y la intervención humana como mecanismo de gobernanza, así como en la evaluación de impacto de protección de datos como herramienta clave, basada en el principio de responsabilidad, en el modelo adoptado por el RGPD.

2.1. La responsabilidad como principio nuclear en el modelo normativo adoptado por el RGPD

En el artículo 5(2) RGPD el principio de responsabilidad constituye un principio nuclear que obliga a cumplir y demostrar el cumplimiento de los otros seis principios enumerados en el artículo 5(1) RGPD, esto es, el principio de licitud, lealtad y transparencia, el principio de limitación de la finalidad, el principio de minimización de datos, el principio de exactitud, el principio de limitación del plazo de conservación y el principio de integridad y confidencialidad. En desarrollo del principio de responsabilidad, el artículo 24 RGPD exige que el responsable del tratamiento tenga en cuenta: «*la naturaleza, el ámbito, el contexto y los fines del tratamiento*, así como *los riesgos de diversa probabilidad y gravedad para los derechos y libertades de las personas físicas* y que, en consecuencia, aplique *medidas técnicas y organizativas apropiadas*», a fin de garantizar

[794] Vid. Apartado 2. La disposición kafkiana: Artículo 22 del RGPD. Razones para su ingreso en la UCI, en Capítulo 2. Toma de decisiones automatizada basada en la elaboración de perfiles en el RGPD: el artículo 22 en la unidad de cuidados intensivos.

[795] Gillis y Simons, «Explanation Justification: GDPR and the Perils of Privacy», 76.

y demostrar el cumplimiento normativo, debiéndose documentar dichas medidas de forma adecuada[796].

Así, dicho principio de responsabilidad construye una gobernanza mixta de la protección de datos. Por un lado, establece unos principios tasados a los que el responsable debe adherirse en el tratamiento de datos personales, si bien y, por otro lado, el modo o forma de cumplimiento de dichos objetivos normativos depende en gran medida de la propia autorregulación que establezca el responsable.

Estamos ante un modelo normativo que, tal y como reflejan distintos artículos y considerandos al referirse a "medidas adecuadas" o "apropiadas", no delimita tanto medidas o niveles de seguridad concretos, como hacía la normativa anterior, sino que parte de la idea de que cada tratamiento concreto precisará de unas medidas de seguridad que se adapten al mismo[797].

No obstante, en algunos casos, el RGPD impone determinadas medidas técnicas u organizativas debido al riesgo que plantea el tratamiento: por ejemplo, en el caso de la toma de decisiones basada únicamente en el tratamiento automatizado en virtud de las excepciones contractuales o basadas en el consentimiento -22(2)(a) y (c) RGPD-, las medidas obligatorias son los derechos del interesado a obtener la intervención humana por parte del responsable del tratamiento, a expresar su punto de vista y a impugnar la decisión -22(3) RGPD-. La obligación de llevar a cabo una EIPD para tratamientos que entrañen un "alto riesgo" para los derechos y libertades de las personas físicas -35 (1) RGPD-, también se impone en este modelo normativo. En otros casos, el RGPD establece o sugiere una serie de medidas que puede adoptar el responsable del tratamiento cuando considere que son adecuadas por la naturaleza, el alcance, el contexto y los fines del tratamiento y sus riesgos para los derechos y libertades de las personas físicas. Por ejemplo, para garantizar la seguridad del tratamiento, el RGPD recomienda, entre otras

[796] Documentar la gestión del riesgo en el tratamiento de datos personales tiene un doble objetivo según la AEPD: en primer lugar, y siendo su objetivo más importante, dar soporte a la ejecución eficaz y eficiente de la gestión del riesgo para los derechos y libertades. En segundo lugar, y supeditado al primero, permitir demostrar que así se ha realizado. Además, la documentación de la gestión del riesgo no es un documento en sí, sino un proceso que se traduce en hechos y que se acredita documentalmente. Agencia Española de Protección de Datos (AEPD), «Gestión del riesgo y evaluación de impacto en tratamientos de datos personales», 12 y 56.

[797] Casanova Asencio, «Mecanismos de prevención del acceso indebido a la historia clínica por parte del personal sanitario y nueva legislación de protección de datos», 5.

cosas, la seudonimización y el cifrado de los datos personales -32(1)(a) RGPD-[798]. Por último, esto no cierra la puerta a otras medidas que, si bien no son requeridas o recomendadas directamente por el RGPD, pueden ser consideradas apropiadas por el responsable del tratamiento para cumplir y demostrar el cumplimiento de la ley, en función del proceso de decisiones que prefiera diseñar. Para evitar la prohibición del artículo 22, apartado 1, el responsable del tratamiento puede aumentar significativamente el nivel de intervención humana de modo que el modelo ya no sea un proceso basado únicamente en el tratamiento automatizado, generalmente prohibido por este artículo[799].

En cierto sentido, y dentro de los límites que marca la propia normativa, no puede imponerse al responsable del tratamiento una forma concreta de cumplir con el RGPD. Sin embargo, esto no relaja los requisitos del mismo para llevar a cabo un tratamiento justo, lícito y transparente y para garantizar el ejercicio de los derechos individuales de los interesados. El RGPD obliga a una gestión integral de los riesgos en el tratamiento de datos, lo cual no quiere decir que la realización de dicha gestión esté en todo caso sometida a las mismas obligaciones jurídicas[800]. Según la AEPD, los mayores perjuicios no se originarán de tratamientos de alto riesgo cuando éstos se encuentren bien gestionados, sino de los tratamientos mal gestionados, en los que se ignoran las amenazas y la gravedad de sus consecuencias[801]. Además, la gestión de este riesgo a través de la aplicación de las medidas oportunas y adecuadas al tratamiento de datos, exige no solo demostrar la conformidad de las actividades sino también la *eficacia* de las medidas aplicadas, conforme al considerando 74 RGPD[802].

---

[798] También se hace mención expresa a la seudonimización como medida apropiada para el tratamiento con fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos en el artículo 89 RGPD.

[799] Así lo contempla Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 33.

[800] Muestra de ello es la EIPD. La gestión del riesgo y la evaluación de impacto para la protección de datos son procesos íntimamente vinculados, no obstante, mientras que la gestión del riesgo es obligatoria para todo tratamiento, las obligaciones concretas que se establecen para la EIPD son obligatorias, exclusivamente, para tratamientos de alto riesgo. Agencia Española de Protección de Datos (AEPD), «Gestión del riesgo y evaluación de impacto en tratamientos de datos personales», 159.

[801] Agencia Española de Protección de Datos (AEPD), 11.

[802] Esta necesidad de demostrar la eficacia de las medidas nos acerca, como veremos más adelante, al enfoque basado en la evidencia que se ha demostrado necesario para asegurar una intervención humana significativa.

Este sistema de gestión de riesgos basado en la responsabilidad como principio nuclear del RGPD se complementa, a su vez, con las posibilidades de vigilancia y rendición de cuentas dadas por la normativa a las personas interesadas y a las autoridades, respectivamente[803]. Por un lado, con el reconocimiento de una serie de derechos individuales que permiten a las personas interesadas, dentro de los límites del ejercicio de los mismos, examinar dicho sistema de gestión de riesgos respecto de -y exclusivamente de- sus datos personales[804], pudiendo incluso exigir la rectificación individualmente y de forma directa al responsable del tratamiento en determinados casos[805]. Por otro lado, con un sistema de vigilancia pública que recae fundamentalmente sobre la figura de las autoridades de control, cuyas competencias, funciones y poderes se establecen en el capítulo VI del RGPD, y que trata de garantizar la rendición de cuentas ante posibles infracciones normativas a través de posibilitar el acceso a recursos ante autoridades públicas -sobre todo de las de control, pero sin cerrar la puerta a otras autoridades que pudieren ser competentes y en particular autoridades judiciales- y sanciones en el capítulo VIII RGPD.

En líneas generales, esta es la forma en la que el principio de responsabilidad o *accountability* se despliega en el RGPD como un proceso activo de generación de conocimiento -documentando el tipo de tratamiento de datos personales, sus riesgos y medidas adoptadas- destinado a ser escrutado -por interesados y autoridades- para restablecer la relación actor-foro haciendo más factible la evaluación de la conducta de

[803] Debe destacarse la importancia del delegado de protección de datos en este sistema de vigilancia y rendición de cuentas para los casos en los que su designación es obligatoria en virtud del art. 37(1) RGPD. Corresponde al delegado tanto la comunicación con los interesados a efectos del ejercicio de sus derechos individuales -art. 38(4) RGPD-, como la cooperación y actuación como punto de contacto con la autoridad de control pertinente -art. 38(1)(d) y (e) RGPD-.

[804] Así hemos podido ver el alcance de los derechos de información y acceso en el marco de la toma de decisiones automatizada basada en la elaboración de perfiles, operando la transparencia como un requisito instrumental de la rendición de cuentas establecida por el RGPD en virtud del principio de responsabilidad. Vid. Apartado 2. Derecho a la información en la toma de decisiones automatizada, en Capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica.

[805] También en el anterior capítulo se ha destacado la permeabilidad de la toma de decisiones automatizada en base a las posibilidades de impugnación del tratamiento de datos personales que ofrece el RGPD. Vid. Apartado 3. Derecho a impugnar las decisiones automatizadas, en Capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica.

aquél -permitiendo evaluar el tratamiento que el responsable (actor) realiza de los datos personales de los interesados (foro)-.

2.2. Evaluación de impacto de protección de datos: la herramienta basada en la responsabilidad para tratamientos de alto riesgo

En el anterior apartado se ha mencionado que, entre otras, la EIPD es una de las medidas organizativas impuestas al responsable del tratamiento en función de riesgo por el RGPD. En particular, cuanto el tratamiento de datos personales entrañe un alto riesgo para los derechos y libertades de las personas físicas.

La interpretación hasta ahora propuesta sobre el artículo 22 y el énfasis sobre la olvidada intervención humana no es la panacea, por razones que ya se han ido recalcando. Primero, porque las limitaciones propias del RGPD como instrumento regulatorio para la toma de decisiones automatizada resultan evidentes, en tanto que dichas disposiciones no establecen obligaciones para el diseño y desarrollo de los sistemas. Segundo, porque en el RGPD no puede entenderse una disposición al margen del -complejo- ecosistema regulatorio que establece. El enfoque aquí propuesto pone de manifiesto esta última. En particular, se propone analizar las obligaciones que el Reglamento establece para los responsables del tratamiento en la evaluación de impacto de protección de datos (EIPD) respecto de los mecanismos de intervención humana para la toma de decisiones automatizada. Son varios motivos los que hacen razonable esta propuesta, atendamos aquí al primero de ellos: el ámbito de aplicación de la EIPD.

El objeto de estudio de esta investigación -toma de decisiones automatizada basada en la elaboración de perfiles- entra dentro del ámbito de aplicación de la EIPD en prácticamente todas sus manifestaciones posibles.

Una vez más, la EIPD debe realizarse con carácter obligatorio para todo tratamiento que probablemente entrañe un alto riesgo para los derechos y libertades de las personas[806]. Además de los criterios establecidos en el articulado del RGPD, para determinar si el tratamiento entraña esta clase de riesgo, el CEPD considera que el responsable del

[806] Artículo 35(1) RGPD: *Cuando sea probable que un tipo de tratamiento, en particular si utiliza nuevas tecnologías, por su naturaleza, alcance, contexto o fines, 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*.

tratamiento debe comprobar los criterios establecidos en los considerandos 71, 75 y 91 del RGPD, en las Directrices del GT29 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[807] y la lista de operaciones de tratamiento que pueden entrañar «un alto riesgo» adoptada a nivel nacional en virtud del artículo 35, apartado 4, que la AEPD ha establecido en su Guía para la gestión del riesgo y evaluación de impacto en tratamientos de datos personales de 2021[808]. En sucesivas ocasiones, el CEPD ha reiterado que el responsable debe considerar obligatoria la realización de la EIPD cuando se cumplan dos de los criterios establecidos, si bien, el cumplimiento de uno solo de ellos puede, en algunos casos, llevar a la misma obligación[809].

[807] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «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», 10 y ss.

[808] Agencia Española de Protección de Datos (AEPD), «Gestión del riesgo y evaluación de impacto en tratamientos de datos personales», 136 y ss.

[809] Comité Europeo de Protección de Datos (CEPD), «Opinion 12/2019 on the draft list of the com supervisory authority of Spain petent regarding the processing operations protection exempt from the requirement of a data impact assessment (Article 35 (5) GDPR)», 4.

Tabla 5. Guía AEPD 2021, p. 134

OBLIGACIÓN DE REALIZAR LA EIPD
"Cuando sea probable que un tipo de tratamiento, en particular si utiliza nuevas tecnologías, por su naturaleza, alcance, contexto o fines, entrañe un alto riesgo para los derechos y libertades de las personas físicas"[114]
Está dentro de alguno de los supuestos establecidos en el artículo 35.3 del RGPD
Existe una norma especial que exige una EIPD para el tratamiento.
Cuando el tratamiento corresponde con alguno de los ejemplos de obligación enumerados en las Directrices WP248.
Cuando el tratamiento cumple al menos dos de las condiciones de las enumeradas en las Directrices WP248 para realizar una EIPD.
Cuando el tratamiento cumpla con dos o más criterios de las *Listas de tipos de tratamientos de datos que requieren evaluación de impacto relativa a protección de datos (art 35.4)* publicada por la AEPD.
Cuando se haya apreciado un alto riesgo teniendo en cuenta los supuestos enumerados en el artículo 28.2 de la LOPDGDD.
Cuando en alguna de las directrices publicadas por el CEPD, el tratamiento esté identificado como obligado a realizar una EIPD.
El tratamiento se encuentre sujeto a un código de conducta o a un mecanismo de certificación que exijan al responsable la realización de una evaluación de impacto.

Por un lado, los usos del *big data* y la IA son claros candidatos a que dicha evaluación de impacto sea obligatoria[810]. No solo porque el uso de nuevas soluciones tecnológicas está, por sí, entre los criterios definidos, sino porque características habituales de estas soluciones tecnológicas también se encuentran entre dichos criterios, como el tratamiento de datos a gran escala o la asociación o combinación de conjuntos de datos[811].

[810] Cotino, «Riesgos e impactos del big data, la inteligencia artificial y la robótica. Enfoques, modelos y principios de la respuesta del Derecho», 29. Más crítica con el margen dado en este sentido por el RGPD se muestra Soriano: *el propio concepto de "alto riesgo" proporciona a responsables y encargados del tratamiento cierto margen de discrecionalidad para decidir en qué casos deben llevarse a cabo las evaluaciones de impacto*. Soriano Arnanz, «La propuesta de Reglamento de inteligencia artificial de la UE y los sistemas de alto riesgo», 115.

[811] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «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», 11-12. También Considerando 91 RGPD: *Lo anterior debe aplicarse, en particular, a las operaciones de tratamiento a gran escala que persiguen tratar una cantidad considerable de datos personales a nivel regional, nacional o supranacional y que podrían afectar a un gran número de interesados y entrañen probablemente un alto riesgo, por ejemplo, debido a su sensibilidad, cuando, en función del nivel de conocimientos técnicos alcanzado, se haya utilizado una nueva tecnología a gran escala y a otras operaciones de tratamiento que entrañan un alto riesgo para los derechos y libertades de los interesados, en particular cuando estas operaciones hace más difícil para los interesados el ejercicio de sus derechos*.

Por otro lado, al margen de que la propia toma de decisiones automatizada con efectos jurídicos o similares es uno de los criterios a considerar por entrañar un alto riesgo[812], el artículo 35 RGPD no distingue entre toma de decisiones basada únicamente en el tratamiento automatizado y la toma de decisiones con previa intervención humana[813], tal y como recoge también el GT29 en sus directrices sobre decisiones individuales automatizadas y elaboración de perfiles[814]. Así, al contrario que los derechos de información y acceso -13(2)(f), 14(2)(g) y 15(1)(h)- o las salvaguardas del artículo 22, esta herramienta del RGPD aborda, a mi juicio, la toma de decisiones desde una perspectiva más amplia, y garantista con los derechos de las personas interesadas.

En definitiva, los responsables que desplieguen sistemas de toma de decisiones automatizadas a partir de tratamientos masivos de datos con el uso de nuevas tecnologías como sistemas de IA y sobre la base de elaboración de perfiles, adopten o no decisiones basadas únicamente en el tratamiento automatizado, deberán llevar a cabo la EIPD pertinente[815]. Desde luego, ninguna duda cabe cuando el tratamiento automatizado de datos para la toma de decisiones automatizada -en sentido amplio- basada en la

[812] Grupo de Trabajo sobre Protección de Datos del Artículo 29, 10. Y, desde luego, resulta obligatoria cuando dicha toma de decisiones se base en la evaluación sistemática y exhaustiva de aspectos personales de personas físicas, como la elaboración de perfiles -art. 35(3)(a) RGPD-.

[813] El artículo 35(3)(a) suprime el término "únicamente" en lo que respecta a la toma de decisiones: *evaluación sistemática y exhaustiva de aspectos personales de personas físicas que se base en un tratamiento automatizado, como la elaboración de perfiles, y sobre cuya base se tomen decisiones que produzcan efectos jurídicos para las personas físicas o que les afecten significativamente de modo similar*;

[814] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 33. Consideramos que esto significa que el artículo 35, apartado 3, letra a), se aplicará en caso de toma de decisiones, incluida la elaboración de perfiles, con efectos jurídicos o significativamente similares que no estén totalmente automatizadas, así como en caso de decisiones basadas únicamente en el tratamiento automatizado definidas en el artículo 22, apartado 1.

[815] Coincido en este punto con Palma Ortigosa cuando señala que la mayoría de los criterios a los que hace referencia el GT29 pueden estar presentes en el despliegue e implementación de sistemas de toma de decisiones automatizadas, sin que en ninguno de los criterios la plena automatización resulte relevante. Palma Ortigosa, «Régimen jurídico de la toma de decisiones automatizadas y el uso de sistemas de inteligencia artificial en el marco del derecho a la protección de datos personales», 225. Lo cual pondría de manifiesto que en la interpretación que realiza el GT29 del RGPD no hay motivos para considerar de mayor riesgo, per se, las decisiones plenamente automatizadas de las decisiones con intervención humana. No obstante, la AEPD no parece seguir este razonamiento, al evaluar los factores de riesgo que se derivan del fin declarado del tratamiento diferencia entre el riesgo "alto" para decisiones automatizadas sin intervención humana y "medio" para el tratamiento como soporte de las decisiones. Conviene matizar, eso sí, que el perfilado, predicción y evaluación de sujetos están considerados de riesgo "alto", vid. Agencia Española de Protección de Datos (AEPD), «Gestión del riesgo y evaluación de impacto en tratamientos de datos personales», 81.

elaboración de perfiles incluye el tratamiento de categorías especiales de datos personales[816].

Además, y más allá del modelo de evaluación de impacto regulado por el RGPD, tanto la doctrina como las instituciones europeas han considerado fundamental la evaluación de impacto como instrumento regulatorio para abordar los efectos perniciosos de la toma de decisiones algorítmica, en particular para el uso de sistemas de IA[817], e incluso su realización se ha considerado indispensable para garantizar el derecho a la vida privada en virtud del artículo 8 del CEDH[818]. Ello se ha hecho patente en las propuestas para la regulación de la IA que ha ido perfilando la Comisión. En el Libro Blanco sobre IA ya se adelantaba que las herramientas de evaluación de conformidad son necesarias para asegurar la fiabilidad, seguridad y cumplimiento normativo de los sistemas de alto riesgo de IA, en particular el cumplimiento de los requisitos de obligado cumplimiento establecidos por dicho documento, entre ellos, la adecuada supervisión humana de los sistemas. El Supervisor Europeo de Protección de Datos (SEPD, en adelante) lamentó que el Libro Blanco no hubiese hecho mención expresa al modelo del RGPD, aunque consideraba que la EIPD sería en cualquier caso obligatoria para los sistemas de IA de alto riesgo[819]. No obstante, la propuesta de Reglamento AIA sí recoge expresamente la evaluación de impacto, obligando a los usuarios[820] a utilizar la información facilitada por los proveedores para cumplir con la obligación de realizar la EIPD -art. 29(6) AIA-. Desde luego, parece evidente que el uso de sistemas de IA que fueren calificados de alto

[816] Incluso cuando este tratamiento no cumpla con el criterio de utilización de datos a gran escala. Considerando 91 RGPD: *La evaluación de impacto relativa a la protección de datos debe realizarse también en los casos en los que se tratan datos personales para adoptar decisiones relativas a personas físicas concretas a raíz de una evaluación sistemática y exhaustiva de aspectos personales propios de personas físicas, basada en la elaboración de perfiles de dichos datos o a raíz del tratamiento de categorías especiales de datos personales, datos biométricos o datos sobre condenas e infracciones penales o medidas de seguridad conexas*.

[817] Un modelo muy interesante es la evaluación de impacto en los derechos humanos propuesto por Mantelero y Esposito para el desarrollo de sistemas de IA, en Mantelero y Esposito, «An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems». También se recomienda la recopilación de evaluaciones de impacto algorítmicas realizada por Kaminski y Malgieri, en Kaminski y Malgieri, «Algorithmic impact assessments under the GDPR: producing multi-layered explanations», 134 y ss.

[818] Volviendo sobre el caso SyRI, la no realización de una EIPD para la puesta en funcionamiento de SyRI fue uno de los elementos fundamentales a la hora de determinar la vulneración del artículo 8(2) CEDH por parte del Tribunal de Distrito de La Haya. Sentencia del Tribunal de Distrito de La Haya [Rechtbank Den Haag] de 5 de febrero de 2020 (ECLI: NL: RBDHA:2020: 865), *SyRI case*, vid. par., 6.103 y 6.106.

[819] EDPS, «Opinion 4/2020 on the European Commission's White Paper on Artificial Intelligence – A European approach to excellence and trust», 15.

[820] La figura del usuario en el AIA se corresponde con la del responsable del tratamiento en el RGPD.

riesgo en esta normativa sobre inteligencia artificial y que tratasen datos personales, deberían igualmente considerarse de alto riesgo a efectos de la normativa de protección de datos.

En definitiva, hemos venido reiterando que el RGPD combina una serie de derechos individuales, entre los cuales encontramos el artículo 22 -cuya aplicación depende de un enfoque basado en el riesgo, con la producción de determinados efectos[821]-, con un modelo de gobernanza sistémico basado en el principio de responsabilidad que se traduce, entre otros, en la obligación de realizar una EIPD -obligación también establecida con un enfoque basado en el riesgo-. Aquí se ha argumentado cómo la EIPD es tan obligatoria como las disposiciones del artículo 22 para la clase de tratamiento de datos personales adoptada en el objeto de estudio de esta investigación -la toma de decisiones automatizada basada en la elaboración de perfiles que produzca efectos jurídicos o significativos-. Es decir, la EIPD no puede escapar al análisis jurídico de la gobernanza de estos sistemas, y tampoco la intersección que se produce entre esta obligación y las contenidas en el artículo 22 RGPD.

2.2.1. Justificación de la toma de decisiones automatizada en la EIPD

La EIPD es un proceso utilizado para reforzar y demostrar el cumplimiento con el Reglamento[822]. Es, por tanto, un instrumento básico para la rendición de cuentas establecida por el principio nuclear de la responsabilidad: «*que ayudan a los responsables no solo a cumplir los requisitos del RGPD, sino también a demostrar que se han tomado medidas adecuadas para garantizar el cumplimiento del Reglamento*». La necesaria determinación de las medidas necesarias para afrontar los riesgos identificados -art. 35(7)(d) RPGD-, en palabras de la AEPD, obliga al responsable a actuar y tiene una

[821] Vid. Apartado 3.4. Efectos jurídicos o de afectación significativa similar: un enfoque basado en el riesgo, en Capítulo 2. Toma de decisiones automatizada basada en la elaboración de perfiles en el RGPD: el artículo 22 en la unidad de cuidados intensivos.

[822] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «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», 4. A pesar de que el RGPD no impone una metodología concreta para realizar la EIPD, sí establece un contenido mínimo en su artículo 35(7), según Palma Ortigosa con dos objetivos*: El objetivo que se pretende con este contenido mínimo es que, por una lado se logre una cierta estandarización a la hora de llevar a cabo la EIPD por parte de distintas organizaciones y, por otro lado, se consiga una aplicación efectiva de esta medida ya que se fijan los parámetros mínimos que el legislador considera necesarios para conseguir una protección adecuada de los derechos y libertades de los interesados*. Palma Ortigosa, «Régimen jurídico de la toma de decisiones automatizadas y el uso de sistemas de inteligencia artificial en el marco del derecho a la protección de datos personales», 231.

dimensión mayor que un mero formalismo plasmado en un documento sobre el que se puedan realizar cambios mínimos para adaptarlo a cualquier tratamiento[823].

Es necesario recalcar que la gestión de riesgos en el RGPD no está orientada a proteger intereses propios del responsable, sino a la protección de la persona, en su dimensión individual y social, como sujeto de los datos o afectado por el tratamiento[824]. Es decir, y en el ámbito que nos concierne, la evaluación de impacto para reforzar y demostrar el cumplimiento normativo se debe realizar desde la perspectiva de protección de los derechos y libertades de los afectados por la toma de decisiones automatizada.

Con carácter general, la EIPD es un instrumento normativo de autoevaluación[825], cuyo valor reside en conducir a la construcción de mejores sistemas en su conjunto[826], para la protección de los derechos y libertades mencionados y, en segunda instancia, para proporcionar un marco de seguridad jurídica al responsable basado en el enfoque de riesgos y la autorregulación[827]. En este sentido, el legislador es consciente de que el riesgo 0 no existe ni resulta exigible[828], y por ello la razón de ser de la EIPD reside en demostrar cómo esos riesgos son mitigados durante el ciclo de vida del tratamiento de datos personales, cómo el responsable del tratamiento ha construido un sistema de tratamiento de datos que cumple con la normativa.

---

[823] Agencia Española de Protección de Datos (AEPD), «Gestión del riesgo y evaluación de impacto en tratamientos de datos personales», 25.

[824] Agencia Española de Protección de Datos (AEPD), 17.

[825] Mantelero, «AI and Big Data: A blueprint for a human rights, social and ethical impact assessment», 768; Hawath, «Regulating Automated Decision-Making: An Analysis of Control over Processing and Additional Safeguards in Article 22 of the GDPR.», 171. No existe obligación jurídica para su publicación, aunque ésta sea recomendable al menos en parte, vid. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «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», 20. No obstante, dados los requisitos documentales y de registro que se establecen, la EIPD sido considerada una "precursora" de la aplicación normativa por parte de las autoridades de control establecidas de acuerdo al artículo 51 RGPD, en Kaminski y Malgieri, «Algorithmic impact assessments under the GDPR: producing multi-layered explanations», 132.

[826] Edwards y Veale, «Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?», 51.

[827] Agencia Española de Protección de Datos (AEPD), «Gestión del riesgo y evaluación de impacto en tratamientos de datos personales», 151.

[828] En palabra de la AEPD: *Siempre existirá un riesgo inherente o inicial implícito en cualquier tratamiento y, una vez que se hayan aplicado medidas y garantías que lo minimicen, seguirá existiendo un riesgo residual*. Agencia Española de Protección de Datos (AEPD), 20.

En el análisis sobre los límites de la intervención humana como mecanismo de gobernanza contenido en el artículo 22 RGPD, concluía que es necesario un enfoque basado en la evidencia para poder garantizar una intervención humana significativa como medida de salvaguarda para las personas afectadas por la toma de decisiones automatizada. A mi juicio, este enfoque está presente en el modelo normativo de la EIPD dentro del ecosistema de gestión de riesgos del RGPD, ¿en qué sentido?

De acuerdo con el considerando 74 RGPD, el responsable del tratamiento no solo está obligado a aplicar medidas oportunas y eficaces y ha de poder demostrar la conformidad de las actividades de tratamiento, sino que ha de poder demostrar también la eficacia de dichas medidas. Es decir, la demostración del cumplimiento normativo no es un ejercicio simplemente descriptivo del tipo de tratamiento de datos personales, sus riesgos y las medidas adoptadas para su mitigación, sino que es también un proceso de documentación empírico que debe permitir evaluar la eficacia de las medidas aplicadas.

Tanto es así, que el RGPD establece un mecanismo excepcional de consulta a la autoridad de control para los casos en los que las medidas establecidas por un responsable del tratamiento resulten ineficaces a la hora de mitigar los riesgos para esta clase de tratamiento y así se demuestre al realizar la EIPD -art. 36-. Es decir, el legislador expone una preocupación manifiesta porque los responsables puedan no ser capaces de mitigar estos riesgos por sí mismos, ofreciendo una vía de asesoramiento por parte de las autoridades de control accesible a todo responsable de tratamiento, siempre y cuando el responsable ya haya realizado la EIPD -art. 36(3)-, para facilitar la mitigación efectiva de los riesgos y el cumplimiento normativo.

Además, en aplicación del carácter continuo de la EIPD[829], la gestión de riesgos a la que obliga el Reglamento incluye el seguimiento y verificación de la eficacia de las medidas adoptadas, realizando un proceso de revisión y reevaluación de las medidas cuando resulte necesario, durante el ciclo de vida completo del tratamiento de datos personales a fin de garantizar que la eficacia de dichas medidas se mantiene, obteniendo los resultados esperados y sin que se produzcan alteraciones en la naturaleza, ámbito, contexto y fines

[829] El art. 35(11) RGPD impone su revisión: *al menos cuando exista un cambio del riesgo que representen las operaciones de tratamiento*.

del tratamiento[830]. En definitiva, las medidas adoptadas deben demostrarse eficaces no solo en el momento anterior al tratamiento, cuando el RGPD impone la realización de la EIPD -art. 35(1)-, sino durante todo el ciclo vital del tratamiento.

También en esta investigación he tenido ocasión de señalar que el enfoque sobre los derechos individuales en la regulación de la toma de decisiones automatizada tiene limitaciones -no solo en lo que respecta a la intervención humana-: ¿qué relevancia tiene observar la "significancia" de la intervención humana desde una perspectiva individual? ¿aporta información relevante el hecho de que se nos faciliten explicaciones individuales sobre la lógica algorítmica? ¿pueden realmente conocerse e impugnarse los efectos discriminatorios de una decisión desde una perspectiva individual?

En el seno del debate doctrinal sobre la extensión de los derechos de información y acceso y el derecho a una explicación[831], encontramos varias publicaciones que son plenamente conscientes de esta limitación y aportan soluciones y argumentos de peso tomando la responsabilidad algorítmica como piedra angular de la regulación de la toma de decisiones automatizada en el RGPD y, en particular, las obligaciones para demostrar el cumplimiento normativo contenidas en la EIPD.

Entre otras aportaciones en este sentido, en varias publicaciones se ha señalado la conveniencia de abandonar la búsqueda de explicaciones sobre la lógica automatizada, por la justificación de la toma de decisiones automatizada[832]. Proveer una explicación sobre una decisión, informando a la persona sobre el resultado o la decisión y sobre las premisas, predicciones o inferencias subyacentes que han conducido a ella, no implica que dicha decisión esté debidamente justificada[833]. Mientras que el objetivo de una explicación es hacer posible que un humano (diseñador, usuario, persona afectada, etc.)

[830] Agencia Española de Protección de Datos (AEPD), «Gestión del riesgo y evaluación de impacto en tratamientos de datos personales», 22 y 50.

[831] Sobre el alcance de éstos, vid. Apartado 2.2. Derecho a la información para las decisiones basadas únicamente en el tratamiento automatizado, ¿derecho a una explicación?, en Capítulo 3. Tres pilares sobre los que interpretar y hacer efectiva la regulación de la toma de decisiones automatizada en el RGPD: derecho a la intervención humana, derecho a la información y derecho a impugnar la decisión. Una propuesta terapéutica.

[832] Entre otras, Gillis y Simons, «Explanation Justification: GDPR and the Perils of Privacy»; Henin y Le Métayer, «A framework to contest and justify algorithmic decisions»; Malgieri, «"Just" Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation».

[833] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI», 12.

pueda entender un resultado algorítmico o el conjunto del sistema, el objetivo de una justificación es convencer de que la decisión es apropiada, por ello, mientras que las explicaciones tienen un carácter descriptivo e intrínseco, las justificaciones son normativas y extrínsecas, en el sentido de que dependen de una referencia según la cual su validez puede evaluarse[834].

En realidad, estos términos no son tan antagónicos como se presentan aquí, justificar requiere a fin de cuentas de una explicación -en el sentido de revelar la información a justificar- que puede manifestarse en una u otra forma según se requiera por el ordenamiento jurídico -a un individuo si hablamos del ejercicio de derechos de información o en una EIPD para el cumplimiento de las obligaciones del responsable del tratamiento en el artículo 35 RGPD-.

En esta línea, Gillis y Simons lamentan que las justificaciones que los derechos de información y acceso exigen a los responsables del tratamiento de cara a los individuos, se limitan a la lógica general del modelo algorítmico y sus consecuencias previstas, y no tanto al conjunto de decisiones que toma una institución sobre el diseño de un modelo de aprendizaje automático y su integración en su procedimiento de toma de decisiones[835]. No obstante, son mucho más optimistas hacia la clase de justificaciones que las evaluaciones de impacto introducen en el modelo de gobernanza sobre la toma de decisiones automatizada. La clase de explicación requerida para la justificación institucional a menudo en la EIPD no será la explicación técnica de la lógica de los modelos de aprendizaje automático a individuos aislados exigida por los derechos de información y acceso[836].

También en este sentido, Malgieri trata de desplazar el foco sobre el derecho a una explicación hacia un marco más amplio de justificación normativa de la toma de decisiones automatizada. Este marco, que vendría introducido por el principio de responsabilidad, impone reclamar procesos de decisión algorítmica que demuestren respetar el núcleo de la protección de datos que en el RGPD se expresa en sus principios

[834] Henin y Le Métayer, «A framework to contest and justify algorithmic decisions», 3.

[835] Gillis y Simons, «Explanation Justification: GDPR and the Perils of Privacy», 87.

[836] Gillis y Simons, 96-97.

recogidos en el artículo 5 del mismo[837]. En realidad, no parece que sea necesario hacer grandes saltos argumentativos para sostener que el principio de responsabilidad en el RGPD reclama del responsable del tratamiento una justificación normativa y que, por ende, la materialización de este principio en la EIPD reclama igualmente una justificación para tratamientos de datos de alto riesgo en los términos definidos en el artículo 35. El RGPD coloca la carga de la prueba en el responsable del tratamiento, de forma que podemos considerar que las decisiones automatizadas en su conjunto son ilegales por defecto, salvo que el responsable del tratamiento las justifique mediante un proceso de justificación válido -en el caso de los tratamientos de alto riesgo este proceso debe adoptar la forma de EIPD-[838].

La evaluación a realizar en este proceso debe incluir, entre otros: «*las medidas previstas para afrontar los riesgos, incluidas garantías, medidas de seguridad y mecanismos que garanticen la protección de datos personales, y a demostrar la conformidad con el presente Reglamento, teniendo en cuenta los derechos e intereses legítimos de los interesados y de otras personas afectadas*» -art. 35(7)(d) RGPD-. La toma de decisiones automatizada basada en la elaboración de perfiles que produce efectos jurídicos o significativos contiene riesgos que el responsable debe identificar en función de las operaciones de tratamiento previstas y de los fines del tratamiento. Una vez identificadas las operaciones de tratamiento y su finalidad, evaluada su necesidad y proporcionalidad e identificados los riesgos -arts. 35(7)(a),(b) y (c) RGPD-, el responsable del tratamiento debe determinar las medidas adoptadas para demostrar el cumplimiento de la normativa[839]. Y, por último, el responsable debe demostrar la eficacia de dichas medidas para garantizar la protección de datos personales teniendo en cuenta los derechos y libertades de las personas que van a ser objeto de la toma de decisiones automatizada.

[837] Malgieri, «"Just" Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation», 22.

[838] Malgieri, 25.

[839] Algunas de estas medidas vienen impuestas por el RGPD, como la introducción de intervención humana significativa previa a la producción de efectos para la toma de decisiones no basada únicamente en el tratamiento automatizado -art. 22(1) RGPD-, o el derecho a obtener intervención humana por parte del responsable, a expresar su punto de vista y a impugnar la decisión para decisiones basadas únicamente en el tratamiento automatizado -art. 22(3) RGPD-. Lo cual no obsta para que el responsable pueda adoptar, además, otras medidas no expresamente previstas y que puedan resultar eficaces para mitigar los riesgos identificados.

En definitiva, para aquellos modelos de toma de decisiones automatizada que entren en el ámbito de aplicación del artículo 35, siendo obligatorio para el responsable llevar a cabo una EIPD, el RGPD está demandando una justificación de las medidas adoptadas para la toma de decisiones automatizada lícita, leal, transparente, exacta, entre otros, por parte del responsable. Bajo el prisma de los argumentos antes aportados, estas justificaciones deben incluir la evaluación de la eficacia de las medidas adoptadas durante todo el ciclo de vida del tratamiento de datos para la toma de decisiones automatizada basada en la elaboración de perfiles.

Esta clase de justificación se demanda independientemente de que la prohibición contenida en el artículo 22(1) RGPD -o 22(4)- se justifique por la introducción de intervención humana significativa previa a la producción de efectos jurídicos o significativos para el interesado, o bien, se justifique por medio de una de las excepciones contenidas en el artículo 22 y con la adopción de las salvaguardas que prevé el RGPD. No obstante, como ya hemos visto, hay dos figuras distintas de intervención humana exigidas por el RGPD en función de si se trata de decisiones basadas únicamente en el tratamiento automatizado o de decisiones en las que el sistema algorítmico es un apoyo a la toma de decisiones, por tanto, la forma en la que la intervención humana se debe justificar para la toma de decisiones automatizada varía igualmente. No obstante, esta perspectiva desde la que podemos entender cómo la intervención humana como mecanismo de gobernanza de la toma de decisiones automatizada se interrelaciona con la obligación de justificar la misma en las obligaciones derivadas de la EIPD se encuentra prácticamente inexplorada.

2.2.2. La intervención humana como medida organizativa para mitigar los riesgos de la toma de decisiones automatizada en la EIPD

Parte de la doctrina ya había identificado la EIPD como una herramienta útil a la hora de determinar el grado de humanización realmente existente en la toma de decisiones automatizada y si dichas decisiones se basan o no únicamente en el tratamiento automatizado[840]. Particular importancia adquiere, a mi juicio[841], analizar el vínculo que,

[840] Veale y Edwards, «Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling»; también en este sentido, Palma Ortigosa, «Automated Decision-Making in the GDPR. Algorithms in the Scope of the Data Protection».

[841] Este vínculo e intersección entre la regulación de la toma de decisiones automatizada y las obligaciones del EIPD sobre el principio de responsabilidad ya ha sido explorado en otras investigaciones aunque no en

en base al principio nuclear de responsabilidad del RGPD, pueda existir entre la intervención humana como mecanismo de gobernanza –con fundamento en dicho principio conforme al análisis realizado anteriormente– y la EIPD como materialización normativa más importante de este principio[842].

Si atendemos a la prohibición contenida en el artículo 22(1) RGPD -y 22(4)- de tomar decisiones basadas únicamente en el tratamiento automatizado que produzcan efectos jurídicos o significativos, conforme al análisis realizado anteriormente podemos concluir que hay dos formas de cumplir con dicha prohibición[843]. O bien se acude a alguna de las excepciones -y salvaguardas- contenidas en el artículo 22 RGPD -y se justifica debidamente-, o bien el responsable de tratamiento puede: «*diseñar un «modelo» de decisiones basadas en la elaboración de perfiles, aumentando significativamente el nivel de intervención humana de forma que el modelo ya no sea un proceso de toma de decisiones totalmente automatizadas*»[844].

En cada caso, el responsable deberá justificar cómo sirve a los objetivos normativos, cómo resulta significativa la intervención y, por ende, por qué es eficaz y cómo se demuestra como tal para mitigar los riesgos del tratamiento en la toma de decisiones automatizada.

En el primer caso, para la toma de decisiones basada únicamente en el tratamiento automatizado, el principio de licitud exige en primer término poder justificar cuál de las excepciones contenidas en el artículo 22 RGPD ha sido aplicada por el responsable del tratamiento. Posteriormente, deberá justificarse la forma en la que se introduce la intervención humana *out of the loop*, con posterioridad a la producción de efectos jurídicos o significativos para la persona interesada. Esta justificación deberá realizarse

detalle sobre los mecanismos de intervención humana. Vid. Kaminski y Malgieri, «Multi-Layered Explanations from Algorithmic Impact Assessments in the GDPR»; Janssen, «An approach for a fundamental rights impact assessment to automated decision-making».

[842] Sin olvidar el deber de protección de datos por diseño, Martínez, «Cuestiones de ética jurídica al abordar proyectos de Big Data. El contexto del Reglamento general de protección de datos», 160.

[843] En cualquiera de los casos, el GT29 parece sugerir que la EIPD requeriría evaluar el grado de participación humana en el modelo de toma de decisiones: *Como parte de la EIPD, el responsable del tratamiento debe identificar y registrar el grado de participación humana en el proceso de toma de decisiones y en qué punto se produce esta*. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679», 23.

[844] Grupo de Trabajo sobre Protección de Datos del Artículo 29, 33.

de todas las medidas -de carácter obligatorio o no- que el responsable adopte para la toma de decisiones basada únicamente en el tratamiento, y no únicamente de la intervención humana como medida organizativa de salvaguarda de los derechos y libertades de las personas interesadas -art. 22(3) RGPD-[845].

Conforme al análisis realizado en esta investigación, este mecanismo de gobernanza se fundamenta en la preocupación por la pérdida de control por parte de los interesados en las decisiones que les afectan de forma directa; la intervención humana en el apartado 22(3) como medida organizativa se encuentra vinculada de forma instrumental con los derechos del interesado a una explicación, a expresar su punto de vista e impugnar la decisión[846]. En este sentido, la eficacia de esta medida y su justificación en la EIPD estarán igualmente vinculadas a esta relación instrumental (¿qué clase de información no tenida en cuenta por el tratamiento automático se recaba a partir de la intervención humana?; ¿cómo se facilita la impugnación de las decisiones automatizadas a partir de la intervención humana?; ¿en qué medida la intervención humana rectifica las decisiones adoptadas de forma totalmente automatizada?). A estos efectos, puede resultar muy útil para demostrar el cumplimiento normativo recabar la opinión de los propios interesados, conforme a las disposiciones sobre la EIPD del propio Reglamento[847].

No obstante, si se diseña un sistema de apoyo a la toma de decisiones, este modelo debe incluir una intervención humana significativa previa -*in the loop*- a la producción de un efecto jurídico o similar en el interesado para cumplir con el RGPD y, por ende, el responsable del tratamiento deberá demostrar igualmente que dicha intervención se produce en estos términos en la EIPD, puesto que constituye una medida organizativa

[845] Por ejemplo, anteriormente hemos visto que para las decisiones basadas únicamente en el tratamiento automatizado el responsable debe facilitar en todo caso los derechos de información y acceso relativos a la aportación de información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas de dicho tratamiento para el interesado (arts. 13(2)(f), 14(2)(g) y 15(1)(h) RGPD). Para demostrar el cumplimiento del principio de transparencia, es necesario que el responsable justifique en el RGPD las medidas adoptadas para el cumplimiento de dichas obligaciones. Muy útil en este sentido el test de justificación propuesto por Malgieri en forma de preguntas para el responsable, vid. Malgieri, «"Just" Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation».

[846] Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems», 1; Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 87; Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations», 22.

[847] Art. 35(9) RGPD: *Cuando proceda, el responsable recabará la opinión de los interesados o de sus representantes en relación con el tratamiento previsto, sin perjuicio de la protección de intereses públicos o comerciales o de la seguridad de las operaciones de tratamiento*.

indispensable -y de obligado cumplimiento- para evitar los riesgos que puedan identificarse en el contexto particular de la toma de decisiones basada en la elaboración de perfiles[848]. En este caso, también conforme al análisis previo, la intervención humana como mecanismo de gobernanza se fundamenta en la preocupación del legislador por la pérdida de control y capacidad para explicar sus propias decisiones por parte de quienes las toman, incurriendo en una dejación de la responsabilidad por el responsable del tratamiento.

Como sabemos, el RGPD no impone un determinado modelo para la toma de decisiones apoyada en sistemas automatizados de elaboración de perfiles, sino que para demostrar el cumplimiento con el artículo 22(1), la intervención humana debe cumplir un requisito temporal -tener lugar de forma previa a la producción de efectos jurídicos o significativos- y un requisito cualitativo -intervención humana significativo-. Incluso el responsable del tratamiento puede decidir incluir un agente humano en la toma de decisiones, *in the loop*, como medida/garantía para la gestión del riesgo en el tratamiento -aunque tuviese la habilitación normativa para tomar dichas decisiones de forma plenamente automatizada-[849].

Ahora, cuando se trata de demostrar la eficacia de la intervención humana como medida organizativa para la reducción del riesgo del tratamiento de datos personales -esquivando de forma legítima la prohibición del artículo 22 RGPD-, se pone de manifiesto, una vez más, que los responsables del tratamiento necesitan mayor seguridad jurídica a la hora de determinar qué puede entenderse por intervención humana significativa -y poder justificarlo debidamente en la EIPD-. En el siguiente apartado trato de aportar algunos

[848] Dado que a los sistemas de toma de decisiones no basados únicamente en el tratamiento automatizado no son aplicables salvaguardas como el derecho a una explicación, o el derecho a expresar su punto de vista e impugnar la decisión -sobre el resto de salvaguardas sí aplicables a esta clase de tratamiento entraremos en el siguiente apartado-, la intervención humana como componente esencial de la toma de decisiones y, por ende, la evaluación sobre su cumplimiento conforme al RGPD adquiere una importancia determinante en la mitigación de los riesgos que esta clase de tratamiento pueda tener sobre los interesados.

[849] Así recoge la AEPD que ese control humano puede establecerse como una medida que actúa en la propia definición de la naturaleza, ámbito, contexto o fines del tratamiento, en las siguientes formas: *Reemplazar tratamientos automatizados por tratamientos manuales que incorporen procedimientos de supervisión y control; Llevar a cabo la supervisión humana de las decisiones automatizadas; Utilizar personal especialmente cualificado en determinadas fases del tratamiento, especialmente en su supervisión*. Agencia Española de Protección de Datos (AEPD), «Gestión del riesgo y evaluación de impacto en tratamientos de datos personales», 104.

criterios claros de los que partir para el diseño de procesos de toma de decisiones con intervención humana significativa y demostrable.

2.2.3. El diseño de procesos de toma de decisiones con intervención humana significativa y demostrable.

De lo desarrollado hasta el momento en relación con las disposiciones relativas a la realización de EIPD por el responsable del tratamiento, podemos entender que un responsable del tratamiento que pretenda implementar un sistema de toma de decisiones automatizado basado en la elaboración de perfiles -objeto de estudio en esta investigación- habrá de cumplir con la obligación de realizar una EIPD conforme al artículo 35 RGPD[850]. Que la realización de la EIPD requiere justificar el cumplimiento normativo del proceso de toma de decisiones, incluyendo las medidas adoptadas para abordar los riesgos de este tratamiento de datos para los derechos y libertades de las personas interesadas, así como la eficacia de estas medidas. Entre las medidas organizativas a adoptar de forma obligatoria, y cuya eficacia debe probarse, encontramos la intervención humana como componente esencial para la toma de decisiones no basada únicamente en el tratamiento automatizado -legítimas conforme a 22(1) RGPD-, y la intervención humana como medida de salvaguarda para decisiones basadas únicamente en el tratamiento automatizado -prohibidas con carácter general por 22(1) RGPD-[851]. A su vez, el RGPD exige una intervención humana significativa, ¿cómo pueden los responsables del tratamiento diseñar procesos de toma de decisiones con intervención humana significativa y demostrable?

Una de las principales críticas relacionadas con la EIPD es que el texto del Reglamento resulta más bien vago en lo que respecta a la metodología concreta que debe aplicarse,

[850] Salvo contadas excepciones. Si el responsable considerase que la toma de decisiones automatizada que pretende implementar no entraña un alto riesgo, debe justificar y documentar los motivos por los que no se realiza una EIPD e incluir/registrar las opiniones del delegado de protección de datos. En caso de duda, la recomendación es que se realice la EIPD: *En los casos en los que no esté claro si se requiere una EIPD, el GT29 recomienda realizar una, ya que esta evaluación representa un instrumento práctico para ayudar a los responsables del tratamiento a cumplir la legislación de protección de datos*. Grupo de Trabajo sobre Protección de Datos del Artículo 29, «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», 9 y 13.

[851] Dado el carácter instrumental de este tipo de intervención humana respecto del resto de medidas de salvaguarda para la toma de decisiones totalmente automatizada -y la menor importancia que adquiere en comparación con la intervención humana como componente esencial de la toma de decisiones automatizada-, no se desarrollarán en mayor profundidad los criterios ya expuestos en el apartado anterior a este respecto.

limitándose a establecer los contenidos mínimos que debe contener[852]. Aunque las autoridades de control y el CEPD han paliado parcialmente esta problemática en los últimos años a través de la publicación de guías y directrices, el análisis de la intervención humana es residual en este sentido, a pesar de la importancia destacada que adquiere en el marco de la toma de decisiones automatizada[853].

En el análisis previo sobre la intervención humana significativa, primero centrado en las directrices del GT29 y después con la interpretación teleológica sobre los fundamentos del artículo 22 RGPD y su precedente, se han extraído una serie de criterios que pretenden servir de base para aportar seguridad jurídica en la interpretación de este término, los cuales pueden resumirse en los siguientes puntos[854]: [1] la intervención humana significativa debe realizarse por una persona con autoridad y competencia para modificar el resultado algorítmico; [2] la intervención humana significativa está fundamentada en que el responsable del tratamiento pueda mantener la responsabilidad sobre el tratamiento automatizado de datos personales y, por ende, teniendo en cuenta todos los datos disponibles, debe contribuir al cumplimiento de los principios de licitud, lealtad y exactitud, entre otros, en la toma de decisiones basada en el tratamiento automatizado de datos; [3] la aplicación rutinaria de los resultados algorítmicos conlleva una dejación de la responsabilidad incompatible con la autonomía y dignidad humana, por ende, la intervención humana significativa debe tener una influencia real sobre el proceso decisorio evitando la aplicación rutinaria de los resultados algorítmicos. Sobre estos criterios hay varias cuestiones que me gustaría resaltar en relación con la demostración de su cumplimiento en la EIPD.

A pesar de que, en coherencia con los principios de protección de datos desde el diseño y por defecto, la EIPD debe realizarse antes del tratamiento -art. 35(1) y (10) RGPD-, ello no quiere decir que no deba actualizarse: «*La actualización de la EIPD a lo largo del*

[852] Gstrein, «European AI Regulation: Brussels Effect versus Human Dignity?», 13.

[853] La propia AEPD declara la necesidad de gestionar los errores técnicos derivados de la implementación automatizada de tratamientos de datos, incluyendo tanto sistemas plenamente automatizados como sistemas de apoyo a la toma de decisiones, reconociendo que la experiencia demuestra que la estimación del impacto que tienen estos fallos y errores en los datos personales no se suele tener en cuenta (pudiendo tener efectos nocivos sobre la calidad de los datos o efectos discriminatorios), en Agencia Española de Protección de Datos (AEPD), «Gestión del riesgo y evaluación de impacto en tratamientos de datos personales», 43.

[854] En lo que se refiere a la intervención humana significativa como componente esencial de la toma de decisiones -art. 22(1) RGPD-, dejando aquí de lado la intervención humana significativa como salvaguarda para la toma de decisiones basada únicamente en el tratamiento automatizado -art. 22 (3) RGPD-.

proyecto de ciclo de vida garantizará que se tenga en cuenta la protección de los datos y la intimidad y propiciará la creación de soluciones que fomenten el cumplimiento»[855]. La EIPD es un proceso continuo, y dicha continuidad adquiere especial relevancia en la demostración del cumplimiento de la prohibición de la toma de decisiones automatizada. Mientras que la autoridad y competencia del agente humano son características más formales-institucionales, la evaluación del resto de criterios requieren de una evaluación más continua que encaja en la propia naturaleza de la EIPD.

Para Noto La Diega, la evaluación sobre si la intervención humana es o no significativa solo podría determinarse caso por caso y, por tanto, la aplicación de las decisiones del 22(1) también debería depender de ese caso por caso[856]. No obstante, hay razones de peso para considerar que el carácter significativo o no de la intervención humana debe determinarse en una evaluación sistémica y continuada del modelo de toma de decisiones. Por un lado, determinar caso por caso el carácter significativo de la intervención humana y, por ende, la aplicación de la prohibición del 22(1) RGPD provocaría una inseguridad jurídica inasumible tanto para responsables del tratamiento como para los interesados[857]. Por otro lado, es insostenible observar en el caso por caso si el agente humano realiza una aplicación rutinaria de los resultados algorítmicos, es decir, no puede determinarse si el modelo de decisiones está afectado por un sesgo de automatización por el hecho de que el agente humano haya seguido o no una decisión particular. Es necesario realizar una

[855] Grupo de Trabajo sobre Protección de Datos del Artículo 29, «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», 16.

[856] Noto La Diega, «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information», 19.

[857] Por supuesto que en el caso concreto se podrá determinar que la intervención humana no ha corregido un error del resultado algorítmico o incluso que ha cambiado un resultado algorítmico que resultaba más apropiado en términos normativos, por ello es tan relevante la posibilidad de impugnar tanto la inferencia algorítmica como la decisión final, esté la misma precedida o no de intervención humana. Sin embargo, dichos errores en el caso concreto no pueden llevarnos a la consideración de que la intervención humana en un modelo de decisiones concreto sea o no significativa. Solo una evaluación sistemática puede determinar este aspecto. Un ejemplo sencillo. Pongamos que estamos evaluando la exactitud del sistema sobre un total de 100 decisiones, el humano ha seguido el criterio algorítmico en 80 de las cuales 75 eran correctas -es decir, el humano se tenía que haber apartado en 5 ocasiones y no lo hizo-, mientras que de las 20 veces que se apartó del criterio algorítmico, la máquina había errado en 15 ocasiones -es decir, el humano debía haber seguido el criterio en otras 5-. Por tanto, el sistema antes de la intervención humana sería exacto en 80/100 ocasiones, mientras que tras la intervención lo sería en 90/100 ocasiones, ¿tendría sentido sostener que la intervención humana no ha sido significativa en las 10 ocasiones en las que se equivocó?

evaluación institucional o sistemática de la intervención humana en el proceso de toma de decisiones[858].

Además, esta evaluación institucional de la intervención humana puede ayudar a los responsables del tratamiento de datos a prevenir los efectos secundarios de esos sistemas en la sociedad que van más allá del nivel individual[859] y, por tanto, a comprobar cómo puede influir la intervención humana en la mitigación de dichos efectos[860]. Determinar desde este punto de vista si la intervención humana es significativa puede resultar complejo, como se ha puesto de manifiesto más arriba, en todo contexto normativo complejo encontraremos distintos valores compitiendo entre sí[861], lo cual puede derivar en que la intervención humana derive -por ejemplo- en un sistema menos exacto, pero más justo, o viceversa, ¿es posible resolver esta tensión?

[858] En este punto pueden ser muy útiles los ítems planteados por Cabitza, Campagner y Datteri para sistemas de IA aplicados al ámbito clínico-asistencial, en Cabitza, Campagner, y Datteri, «To Err is (only) Human. Reflections on How to Move from Accuracy to Trust for Medical AI».:

-Cuántas veces los responsables de la toma de decisiones (DM) y el sistema están de acuerdo en un caso (concordancia);

-Cuántas veces los DM creen que el sistema tiene razón (confianza);

-Cuántas veces se demuestra que el sistema tiene razón después de los hechos (precisión);

-Cuántas veces los DMs han cambiado de opinión por el consejo del sistema (impacto en el rendimiento);

-Cuántas veces los DMs han percibido que el sistema es útil (utilidad);

-Cuántas veces los DMs han creído recibir elementos interesantes para tomar sus decisiones (utilidad);

-Cuántas veces creen haber sido más rápidos en su toma de decisiones o, más bien, obstaculizados por el sistema (satisfacción);

-Cuántas veces el resultado del sistema ha facilitado o censurado la discusión con sus colegas o con los pacientes (impacto colaborativo);

-Cuántas veces ha facilitado el aprendizaje o ha aliviado de la "carga" el recuerdo, el razonamiento analógico y la inferencia deductiva (impacto cognitivo):

-Cuántas veces los usuarios creen que una herramienta de este tipo puede haber alimentado el sesgo de confirmación, la medicina defensiva, y otros sesgos (como el sesgo de automatización y la complacencia).

[859] Tamò-Larrieux, «Decision-making by machines: Is the ‘Law of Everything’ enough?», 14-15.

[860] La aplicación de medidas técnicas y organizativas apropiadas para garantizar y demostrar el cumplimiento del RGPD se debe hacer teniendo en cuenta *la naturaleza, el ámbito, el contexto y los fines del tratamiento así como los riesgos de diversa probabilidad y gravedad para los derechos y libertades de las personas físicas* (GT29 2017, 22). Aunque puede considerarse que la evaluación de riesgos sobre los derechos y libertades de las personas físicas en la EIPD (arts. 35(1) y 35(7)(c)), incluyen los derechos fundamentales, el RGPD no incluye enfoques detallados sobre cómo o con qué nivel de detalle deben evaluarse el impacto sobre dichos derechos fundamentales (Janssen 2020, 85).

[861] Brennan-Marquez y Henderson, «Artificial Intelligence and Role-Reversible Judgment», 142.

Parece complicado y el trabajo jurídico por realizar en este campo es amplio[862].

Llegados a este punto, parece importante recalcar las limitaciones que tanto la EIPD como la intervención humana tienen como mecanismos de gobernanza de los sistemas algorítmicos. Tal y como se señalaba anteriormente, la EIPD es un instrumento normativo de autoevaluación con una publicidad muy limitada, cuyo valor reside en conducir a la construcción de mejores sistemas en su conjunto responsabilizando al responsable del tratamiento del cumplimiento del RGPD y su demostración. El rol de la intervención humana debe entenderse en ese conjunto. Del mismo modo que la evaluación de la intervención humana significativa es sólo una de las medidas que deben/pueden adoptarse y justificarse en el diseño del proceso de toma de decisiones automatizada, y que pueden servir no sólo para evitar errores, sesgos y discriminaciones, sino también para legitimar un sistema o incluso respetar la dignidad de un individuo dentro de él[863].

Además, el RGPD no impone un grado y clase concreto de intervención humana[864]. Con lo cual, los responsables del tratamiento cuentan con un amplio abanico de posibilidades para cumplir con la intervención humana como componente esencial de la toma de decisiones impuesta por el Reglamento y demostrar su cumplimiento en la EIPD. La evidencia muestra cómo las interacciones bien diseñadas entre la inteligencia humana, la inteligencia de las máquinas y las medidas organizativas pueden mitigar los efectos discriminatorios[865] y mejorar la exactitud en la toma de decisiones[866]. Coincidiendo plenamente con las palabras de Binns hay un modelo de colaboración humano-máquina

[862] Podría aventurarse que, en el contexto de la elaboración de perfiles y la toma de decisiones automatizada, el RGPD considera una prioridad la mitigación de los efectos discriminatorios, en su considerando 71. Ello podría dar pie a justificar en la EIPD que la intervención humana puede tener un carácter significativo por contribuir a dicha mitigación, aunque lo haga en sacrificio de cierto grado de exactitud.

[863] Kaminski y Malgieri, «Algorithmic impact assessments under the GDPR: producing multi-layered explanations», 133.

[864] Tampoco los tribunales deben imponer su propia visión sobre cómo cumplir con los requisitos del GDPR en esta materia. Así lo expresaba, en primera instancia, la Sentencia del *England and Wales High Court*, de 4 de septiembre de 2019, en el caso *R (on the application of Edward Bridges) v the Chief Constable of South Wales Police*. [2019] EWHC 2341 (Admin), caso núm. CO/4085/2018: *(146) On a complaint about a failure to comply with section 64 Data Protection Act 2018, it is for the Court to decide whether the data controller has discharged that obligation. What is required is compliance itself, i.e. not simply an attempt to comply that falls within a range of reasonable conduct. However, when determining whether the steps taken by the data controller meet the requirements of section 64, the Court will not necessarily substitute its own view for that of the data controller on all matters*.

[865] Berendt y Preibusch, «Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop-and Under the Looking Glass», 149.

[866] Vid. Cabitza, Campagner, y Sconfienza, «Studying human-AI collaboration protocols: the case of the Kasparov's law in radiological double reading».

implícito en el artículo 22(1) RGPD[867] y la EIPD es, a mi juicio, una herramienta adecuada para evaluar, y reevaluar si fuera necesario, el mejor modelo de colaboración posible dentro de los amplios márgenes establecidos en dicha disposición[868].

En resumen, el artículo 35 del RGPD exige a los responsables del tratamiento que demuestren continuamente cómo se introduce la intervención humana para cumplir con el derecho del interesado a no ser sometido a una toma de decisiones individual automatizada. Esto no significa que el Reglamento imponga un modelo específico de toma de decisiones, ni que deba entenderse al margen del resto de medidas y salvaguardas para demostrar que dicha intervención humana resulta significativa[869].

2.2.4. Necesidad de tender puentes entre el desarrollo y la implementación del modelo: al hilo de la propuesta de Reglamento AIA

En última instancia, para garantizar una intervención humana significativa y demostrable, es necesario referirse a la indisoluble relación existente entre el diseño y desarrollo de un modelo y de cómo esta fase modula y constriñe las posibilidades de supervisión humana en la fase de implementación del modelo algorítmico en los procesos de toma de decisiones. Así, los beneficios previstos al implantar la automatización por desarrolladores de los modelos y quienes deciden utilizarlos -aumento de la eficiencia, mejora de la seguridad, aumento de la flexibilidad de las operaciones, reducción de la carga de trabajo de los operarios, etc.- no siempre se materializan y pueden verse contrarrestados por los costes de rendimiento humano asociados al uso inadaptado de una automatización mal diseñada o inadecuadamente formada[870].

Por ello, resulta indispensable asegurar la utilización de los modelos de IA conforme al fin y uso para el que fueron diseñados y desarrollados, es decir, la configuración del

[867] Binns, «Human Judgment in algorithmic loops: Individual justice and automated decision-making», 9.

[868] Selbst también se muestra optimista sobre el potencial de las evaluaciones de impacto en el diseño de colaboraciones humano-máquina para el uso de sistemas automatizados en la prevención del crimen. Vid. Selbst, «Disparate Impact in Big Data Policing».

[869] Al igual que la intervención humana como medida de salvaguarda -22(3) RGPD- tiene un carácter instrumental respecto de la impugnación de las decisiones por los interesados, y su significancia puede justificarse a partir de cómo sirve a dicho carácter instrumental; para la toma de decisiones no basada únicamente en el tratamiento automatizado, pueden adoptarse medidas alternativas, técnicas u organizativas, que sirvan instrumentalmente para garantizar una intervención humana significativa y demostrable como componente esencial de la toma de decisiones -22(1) RGPD-.

[870] Parasuraman y Manzey, «Complacency and Bias in Human Use of Automation: An Attentional Integration», 381.

elemento humano en el diseño y desarrollo juega un papel clave en la futura implementación[871]. En este sentido, es necesario que el responsable del tratamiento evalúe correctamente las características bajo las que fue desarrollado y validado el sistema, de forma que pueda considerar qué cambios ha de introducir en el entorno existente en el que se implementará para asegurar un funcionamiento adecuado del sistema (y considere también el coste -en términos económicos y humanos- de dichos cambios).

Por supuesto, resulta apropiado para demostrar el cumplimiento normativo asegurar que un modelo es utilizado conforme al fin y uso para el que fue diseñado y justificar este uso en la EIPD. No obstante, esta posibilidad encuentra severas limitaciones. Por un lado, como ya se ha explicado, aunque las disposiciones del RGPD pueden afectar al diseño y desarrollo de modelos algorítmicos cuando en dicha fase se traten datos personales, su impacto es más bien tangencial y, desde luego, no hay disposición alguna que vincule los mecanismos de intervención humana en el RGPD con las fases de diseño y desarrollo de los modelos. Por otro lado, el diseño y desarrollo de determinados modelos algorítmicos puede estar sujeto a normativa sectorial -por ejemplo, la normativa europea de productos sanitarios para los sistemas de IA con fines médicos[872]-, aunque esto no es aplicable a todos los sistemas y, además, la legislación sectorial en cuestión podría no abordar de forma adecuada estos riesgos[873].

Esta laguna en la normativa vigente se pone de manifiesto en las sucesivas propuestas para la regulación europea de los sistemas de IA. Primero con el reconocimiento de la supervisión humana como un requisito de obligado cumplimiento para los sistemas de IA de alto riesgo y, segundo, con el reconocimiento de que la gobernanza de estos sistemas

[871] Jamieson y Goldfarb, «Clinical considerations when applying machine learning to decision-support tasks versus automation», 779.

[872] Con normativa europea de productos sanitarios me refiero al Reglamento (UE) 2017/745 del Parlamento Europeo y del Consejo, de 5 de abril de 2017, sobre los productos sanitarios -MDR, en adelante-, y al Reglamento (UE) 2017/746 del Parlamento Europeo y del Consejo, de 5 de abril de 2017, sobre los productos sanitarios para diagnóstico in vitro.

[873] Siguiendo con el ejemplo anterior, parece que el modelo regulatorio actual de productos sanitarios resulta muy poco satisfactorio en este sentido. Entre las obligaciones particulares para los sistemas informáticos no se incluye ninguna disposición sobre la interfaz humano-máquina -anexo I (17) MDR-.

requiere de garantizar el cumplimiento de distintos requisitos normativos a lo largo de todo el ciclo de vida de los productos y sistemas de IA[874].

¿Cómo se ha traducido esto en la propuesta de Reglamento conocida como "Ley de Inteligencia Artificial" de la Comisión Europea (en adelante propuesta de Reglamento AIA o AIA)?

Aunque ya se ha desarrollado anteriormente en esta investigación, es pertinente recordar que el artículo 14 establece que los sistemas de alto riesgo deberán diseñarse y desarrollarse de modo que puedan ser supervisados de forma efectiva por personas físicas durante su fase de uso -art. 14(1) AIA-. El fundamento de esta supervisión efectiva está en prevenir o reducir al mínimo los riesgos para la salud, la seguridad o los derechos fundamentales, tanto cuando se utiliza un sistema de IA conforme a su finalidad prevista o se le da un uso indebido *razonablemente previsible* -art. 14(2) AIA-[875]. Así, la Comisión obliga al desarrollador a adoptar determinadas medidas para garantizar esta supervisión desde el diseño y desarrollo del modelo, o bien definiéndolas e integrándolas en el sistema de IA -cuando sea técnicamente posible-, o bien definiéndolas para que las lleve a cabo el propio usuario del sistema -art. 14(3) AIA-. Las "cualidades" a partir de las cuales se establece el concepto de "supervisión efectiva" que permita al proveedor cumplir con dicho requisito, también vienen definidas en el articulado; así, las personas a quienes se encomiende la supervisión del sistema deben ser capaces de, entre otros, entender por completo las capacidades y limitaciones del sistema, ser conscientes del sesgo de automatización, interpretar correctamente la información de salida del sistema, desestimar, invalidar o revertir dicha información o interrumpir el sistema accionando un botón específicamente destinado a tal fin -art. 14(4) AIA-.

[874] Sobre estas dos cuestiones, vid. Apartado 3.1. Propuestas para la regulación europea de la inteligencia artificial, en Introducción a la gobernanza y supervisión humana de la toma de decisiones automatizada basada en la elaboración de perfiles, y Apartados 1. Fases de la toma de decisiones y su adecuación al RGPD y 2.2.1. La supervisión humana en las propuestas europeas de regulación de los sistemas de IA de alto riesgo, en Capítulo 1. Marco teórico de la toma de decisiones automatizada basada en la elaboración de perfiles.

[875] Esta previsibilidad se vincula necesariamente con el sistema de gestión de riesgos establecido por la propuesta, otorgando además a la supervisión humana un carácter de salvaguarda última cuando este apartado añade: *en particular cuando dichos riesgos persisten a pesar de aplicar otros requisitos establecidos en el presente capítulo*.

Es importante recalcar que la propuesta de la Comisión Europea no establece obligaciones jurídicas concretas para la supervisión humana en la fase de uso del sistema[876]. No obstante, sí parece que la propuesta, en todo caso, obliga a los usuarios a aplicar las medidas de supervisión humana indicadas por el proveedor y según las instrucciones específicas[877].

De particular relevancia resulta la obligación de utilizar por los usuarios -responsables del tratamiento desde la perspectiva de la normativa de protección de datos- la información facilitada por los proveedores o desarrolladores para el cumplimiento de la EIPD[878]. Las medidas de supervisión humana se incluyen específicamente entre la información que los proveedores deben proporcionar a los usuarios/responsables en virtud del requisito de transparencia, incluidas las medidas técnicas establecidas para facilitar la interpretación de los resultados de los sistemas de IA por parte de los usuarios -art. 13(3)(d) AIA-. Por lo tanto, en virtud de la propuesta de Reglamento AIA, los usuarios/responsables deberán utilizar la información facilitada por los proveedores -que permite a las personas a las que se asigna la supervisión humana por mandato del artículo 22 RGPD comprender las capacidades y limitaciones del sistema- para cumplir, a su vez, con el mandato del artículo 35 del RGPD. A mi modo de ver, esta propuesta ampliará las posibilidades de que los usuarios proporcionen y demuestren una intervención humana significativa como responsables del tratamiento de datos en el RGPD.

Además, un sistema de seguimiento posterior a la comercialización de los sistemas de IA, es decir para la fase de implementación y uso de los mismos -como el diseñado en la propuesta de Reglamento AIA[879]- podría ayudar a identificar y modificar los sistemas de IA de alto riesgo que no puedan ser supervisados eficazmente por personas físicas y, por tanto, no permiten a los responsables del tratamiento cumplir con la intervención humana en virtud del artículo 22 RGPD, por no haber intervención humana significativa posible

[876] No obstante, es posible que para determinados usos de sistemas de IA el usuario del sistema esté obligado por otra norma a introducir un supervisor humano en esta fase bajo determinadas particularidades, en particular, los mecanismos de gobernanza basados en la intervención humana del artículo 22 RGPD para el caso de sistemas de IA que realicen perfiles sobre el tratamiento de datos personales.

[877] Schwemer, Tomada, y Pasini, «Legal AI Systems in the EU's proposed Artificial Intelligence Act», 7.

[878] Art. 29(6) AIA: *Los usuarios de sistemas de IA de alto riesgo utilizarán la información facilitada conforme al artículo 13 para cumplir la obligación de llevar a cabo una evaluación de impacto relativa a la protección de datos que les imponen el artículo 35 del Reglamento (UE) 2016/679 o el artículo 27 de la Directiva (UE) 2016/680, cuando corresponda*.

[879] Título VIII AIA.

dadas las limitaciones introducidas desde el diseño y desarrollo del modelo. Al llevar a cabo una EIPD, los responsables del tratamiento recopilarán datos y pruebas sobre si es posible que la intervención humana contribuya al cumplimiento del RGPD -en los términos arriba planteados- durante el uso de un sistema de IA de alto riesgo. Y, como usuarios en virtud de la propuesta de Reglamento AIA, los responsables del tratamiento podrían compartir esa información recopilada con los desarrolladores para permitirles evaluar el cumplimiento del sistema con el requisito de supervisión humana a ojos del AIA[880], es decir, para demostrar que sus sistemas de IA de alto riesgo pueden ser supervisados eficazmente por personas físicas.

Con sus limitaciones[881], el modelo de regulación para la IA propuesto por la Comisión en combinación con el RGPD podría reforzar la gobernanza y supervisión humana de los procesos de toma de decisiones automatizada basada en la elaboración de perfiles en dos direcciones distintas. Por un lado, el establecimiento de la supervisión humana como un requisito de obligado cumplimiento desde el diseño y desarrollo de los modelos -para los desarrolladores- podrá contribuir de forma efectiva al cumplimiento normativo -por parte de los responsables del tratamiento- y su demostración -a través de la EIPD- de los mecanismos de gobernanza basados en la intervención humana aplicables a la fase de

[880] Art. 61(2) AIA: *El sistema de seguimiento posterior a la comercialización recabará, documentará y analizará de manera activa y sistemática datos pertinentes proporcionados por usuarios o recopilados a través de otras fuentes sobre el funcionamiento de los sistemas de IA de alto riesgo durante toda su vida útil, y permitirá al proveedor evaluar el cumplimiento de los requisitos establecidos en el título III, capítulo 2, por parte de los sistemas de IA.*

[881] Entre otras críticas al AIA, se ha destacado que las personas afectadas por los sistemas de IA no son mencionadas en ninguna ocasión como tampoco posibles mecanismos de garantía o tutela de sus derechos. Por ello, parece entenderse como que la perspectiva de los derechos de los afectados queda completada con la regulación de protección de datos. La permeabilidad de este modelo normativo, entendida como la capacidad de la persona afectada para influir sobre el tratamiento automatizado, es prácticamente inexistente. Entre otros, Cotino et al., «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)»; Veale y Zuiderveen Borgesius, «Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach». También en esta línea el CEPD y SEPD conjuntamente reclamaban la inclusión de derechos para los individuos afectados: *El Reglamento deberá promover formas nuevas , más proactivas y oportunas de informar a los usuarios de los sistemas de IA sobre la situación (de toma de decisiones) en que se encuentra el sistema en cualquier momento, proporcionando una alerta temprana de posibles resultados perjudiciales, de modo que las personas cuyos derechos y libertades puedan verse perjudicados por decisiones autónomas de las máquinas puedan reaccionar o corregir la decisión*. Comité Europeo de Protección de Datos (CEPD) y Supervisor Europeo de Protección de Datos (SEPD), «Dictamen conjunto 5/2021 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)», 22. También proponen la inclusión de derechos individuales para los afectados como un derecho a objetar el uso de estos sistemas, un derecho a impugnar sus decisiones o un a ser informados sobre incidentes graves que puedan tener un impacto significativo sobre sus derechos y libertades -similar al artículo 34 del GDPR-, en Cserne, Ducato, y Živković, «Commentary to the Commission's proposal for the "AI Act" – Response to selected issues», 14.

implementación del sistema -art. 22 RGPD-. Por otro lado, un sistema de seguimiento posterior a la comercialización que promueva la comunicación de los datos recabados en la fase de implementación del sistema -por parte del responsable del tratamiento a través de la EIPD- permitirá al proveedor evaluar el cumplimiento -por parte de los proveedores- de los requisitos de obligado cumplimiento -entre otros, de la supervisión humana- para el desarrollo y sistemas de IA de alto riesgo durante todo el ciclo de vida del sistema.

3. Reflexiones provisionales sobre el capítulo cuarto – tentative thoughts on chapter four

En este apartado se recogen una serie de reflexiones provisionales a modo de cierre de cada capítulo. Aunque algunas de estas reflexiones servirán de apoyo para las conclusiones de esta investigación, el objetivo de este apartado no es exponer dichas conclusiones propiamente, sino resaltar de forma telegráfica algunos aspectos clave resultado del análisis realizado en cada capítulo.

- La intervención humana como mecanismo de gobernanza para la toma de decisiones automatizada basada en la elaboración de perfiles tiene importantes limitaciones. Dichas limitaciones apuntan en dos direcciones: (1) la gobernanza de esta clase de toma de decisiones no puede hacerse depender exclusivamente de la intervención humana; (2) la intervención humana debe responder a objetivos normativos realistas -basados en la evidencia- para que su efectividad resulte demostrable.

- La introducción de la responsabilidad sobre el cumplimiento normativo y su demostración como principio nuclear del RGPD ha consolidado el control colectivo sobre la justificación de los procesos que recopilan, procesan y usan los datos personales, incluyendo los procesos de toma de decisiones automatizada basados en la elaboración de perfiles.

- El principio nuclear de la responsabilidad obliga al responsable del tratamiento a adherirse a los principios del RGPD, aunque lo hace permitiendo un cierto grado de autonomía a la hora de autorregular la identificación de los riesgos y las medidas adoptadas para cumplir con la normativa en el tratamiento de datos personales, siempre que demuestre debidamente dicho cumplimiento.

- Cuando el tratamiento de datos personales conlleva un alto riesgo para los derechos y libertades de las personas físicas, la demostración del cumplimiento normativo exige, necesariamente, la realización de una EIPD. En la práctica totalidad de los supuestos, los procesos de toma de decisiones automatizada basada en elaboración de perfiles que producen efectos jurídicos o significativos incurren en dicho alto riesgo, siendo de aplicación obligada esta medida organizativa para estos procesos de toma de decisiones. A estos efectos, el RGPD no distingue entre decisiones basadas únicamente, o no, en el tratamiento automatizado -35(3)(a)-.

- Para llevar a cabo la EIPD sobre un proceso de toma de decisiones automatizada, el responsable del tratamiento debe, una vez identificadas las operaciones de tratamiento y su finalidad, evaluada su necesidad y proporcionalidad e identificados los riesgos -35(7)(a),(b) y (c)-, determinar las medidas adoptadas para demostrar el cumplimiento de la normativa -35(7)(d)- y, a su vez, demostrar la eficacia de dichas medidas -considerando 74-.

- La intervención humana, analizada como un mecanismo para la toma de decisiones automatizada desde la perspectiva del artículo 22 RGPD, constituye también una medida organizativa desde la óptica de la gestión integral de riesgos y de la realización de la EIPD. Tanto si se trata de decisiones basadas únicamente en el tratamiento automatizado -22(3): *human out of the loop*-, como si no -22(1): *human in the loop*-, el responsable debe justificar cómo se produce la intervención humana en el proceso y cómo ésta resulta significativa.

- El RGPD abre un margen de discrecionalidad amplio a la hora de que el responsable diseñe una intervención humana significativa como componente esencial de la toma de decisiones -22(1)-, siendo la EIPD una herramienta de autoevaluación adecuada para evaluar, y reevaluar si fuera necesario, el mejor modelo de colaboración humano-máquina posible dentro dicho margen.

- Para la consecución de una intervención humana significativa es determinante la forma en la que el sistema ha sido diseñado y desarrollado. En el RGPD no encontramos vínculo normativo alguno entre la fase de diseño y desarrollo de los modelos algorítmicos y los mecanismos de intervención humana establecidos para

la fase de implementación de los mismos. La propuesta de Reglamento AIA podría aportar soluciones a esta laguna.

As a method of recapping each chapter, this section presents a number of tentative thoughts. The research's conclusions will be supported by some of these insights, but the purpose of this section is not to present those conclusions in their entirety. Rather, it aims to emphasize certain key aspects that came out of the analysis done in each chapter in a telegraphic manner.

- Human intervention as a governance mechanism for automated decision-making based on profiling entails important limitations. These limitations point in two directions: (1) the governance of this kind of decision-making cannot be made to rely exclusively on human intervention; (2) human intervention must respond to realistic - evidence-based - policy objectives in order to be demonstrably effective.
- The introduction of accountability for compliance and its demonstration as a core principle in the GDPR has strengthened collective control over the justification of procedures that collect, process and use personal data, including automated decision-making processes based on profiling.
- The core principle of accountability requires the controller to adhere to the principles of the GDPR, while allowing a certain degree of autonomy to self-regulate the identification of risks and the measures taken to comply with the GDPR in the processing of personal data, provided that it duly demonstrates such compliance.
- Where the processing of personal data entails a high risk to the rights and freedoms of natural persons, the demonstration of compliance necessarily requires the performance of a DPIA. For automated decision-making processes based on profiling that produce legal or significant effects, in virtually all cases, such a high risk is incurred, and hence to carry out a DPIA is mandatory for such decision-making processes. On this matter, the GDPR does not distinguish between decisions based solely, or not, on automated processing - 35(3)(a) -.
- In order to carry out a DPIA for an automated decision-making process, once identified the processing operations and their purpose, assessed their necessity and proportionality and identified the risks -35(7)(a), (b) and (c)-, the controller must

determine the measures taken to demonstrate compliance - 35(7)(d) - and, in turn, demonstrate the effectiveness of those measures -Recital 74-.

- Human intervention, analysed before as a governance mechanism for automated decision-making from the perspective of Article 22 GDPR, also constitutes an organisational measure from the perspective of integrated risk management and the implementation of DPIAs. Whether decisions are based solely on automated processing -22(3): human out of the loop- or not -22(1): human in the loop- the controller must justify how human intervention in the process is integrated and how it is meaningful.
- The GDPR opens a wide margin of discretion for the controller to design meaningful human intervention as an essential component of decision-making -22(1) - whereas the DPIA is an appropriate self-assessment tool to evaluate, and re-evaluate if necessary, the best possible model of human-machine collaboration within this margin.
- The way in which the system has been designed and developed is decisive for the implementation of meaningful human intervention. Under the GDPR there is no regulatory link between the design and development phase of algorithmic models and the human intervention mechanisms established for the implementation phase. The proposed AIA Regulation could provide solutions to this gap.

CONCLUSIONES SOBRE LA GOBERNANZA Y SUPERVISIÓN HUMANA DE LA TOMA DE DECISIONES AUTOMATIZADA BASADA EN LA ELABORACIÓN DE PERFILES

-

CONCLUSIONS ON THE GOVERNANCE AND HUMAN OVERSIGHT OF AUTOMATED DECISION MAKING BASED ON PROFILING

CONCLUSIONES SOBRE LA GOBERNANZA Y SUPERVISIÓN HUMANA DE LA TOMA DE DECISIONES AUTOMATIZADA BASADA EN LA ELABORACIÓN DE PERFILES

No quisiera que el marco teórico desarrollado en esta investigación ofrezca una imagen reduccionista de las problemáticas de la toma de decisiones automatizada. En línea con los argumentos de Hildebrandt; reducir el problema de la aplicación de modelos algorítmicos a su naturaleza de caja negra u opaca, a los potenciales sesgos que puede crear o reforzar o a la supervisión humana que pueda establecerse sobre los mismos, puede distraer la atención sobre cuestiones inherentes y fundamentales de estos procesos[882]. Principalmente, los sacrificios inherentes a la introducción de lógicas algorítmicas en distintos procesos de toma de decisiones de nuestro entorno social; automatizando, datificando y regulando bajo parámetros estadísticos, en mayor o menor medida, la actividad humana en la que se decide implementar un sistema de estas características.

En cualquier caso, sí tengo esperanza en que el marco teórico desarrollado sea de utilidad para juristas que pretenden acercarse a la materia, como herramienta para comprender, a grandes rasgos: qué fases integran los sistemas de toma de decisiones automatizada basada en la elaboración de perfiles y la relevancia jurídica de cada una de ellas; las distintas formas en las que el Derecho puede hacer partícipes a los humanos de esos procesos para servir a distintos objetivos normativos; qué son los sesgos, cómo se reproducen en un sistema automatizado y su distinto significado en los planos estadístico, ético y jurídico; las distintas clases de opacidad a las que puede hacerse referencia al hablar del efecto caja negra de los modelos algorítmicos y cómo cada clase de opacidad merece una respuesta jurídica distinta y apropiada al contexto en el que se implementa un sistema.

Esta reflexión inicial sobre el capítulo primero de este trabajo da paso, a continuación, a las conclusiones finales extraídas del análisis de la normativa europea de protección de datos para la gobernanza y supervisión humana de la toma de decisiones automatizada basada en la elaboración de perfiles.

[882] Hildebrandt, «Data-Driven Prediction of Judgment. Law's New Mode of Existence?»

PRIMERA.

Si observamos la regulación de la toma de decisiones automatizada basada en la elaboración de perfiles en el RGPD desde una perspectiva puramente teórica, podemos concluir que, mientras muchos gobiernos y empresas tecnológicas sueñan con eliminar a los seres humanos de los procesos de toma de decisiones, el RGPD parece ir en dirección contraria y muestra en el artículo 22 una clara preferencia política por el *human in the loop*[883]. En ese sentido, la normativa de protección de datos se encuentra plenamente alineada con las iniciativas regulatorias de los sistemas de IA de alto riesgo surgidas en el seno de la UE, y que sin excepciones han situado la supervisión humana como un requerimiento central y de obligado cumplimiento para el desarrollo e implementación de estos sistemas. No obstante, desde una perspectiva más práctica acerca de la influencia del RGPD en el mundo real, en este trabajo se ha constatado un nuevo fracaso de la normativa europea de protección de datos para asegurar el cumplimiento de la regulación de la toma de decisiones automatizada. Esta incapacidad se hace más patente si cabe en lo que se refiere a los mecanismos de gobernanza basados en la intervención humana. Sobre estos mecanismos, con escasa relevancia en la -ya de por sí- poca jurisprudencia que ha interpretado el artículo 22, la doctrina ha centrado el análisis en la intervención humana como medida de salvaguarda de la toma de decisiones basada únicamente en el tratamiento automatizado -22(3): *human out of the loop*-, y no tanto en la intervención humana como componente esencial para la toma de decisiones no basada únicamente en el tratamiento automatizado -22(1): *human in the loop*-.

SEGUNDA.

Al margen de otros factores que puedan explicar el fracaso de esta disposición, resulta obvio que la redacción adolece de la claridad y simplicidad necesaria para cualquier norma jurídica que pretenda ser aplicada de forma efectiva. Sin ánimo de ser exhaustivo, dado que aquí las mejoras podrían ser de muy distinto calado, dependiendo de la clase de solución que se le quisiera dar; en su apartado primero y cuarto debería especificarse que se trata de dos prohibiciones generales de adoptar decisiones basadas únicamente en el tratamiento automatizado -y no de un derecho a interponer por el interesado-; en el apartado primero debería incluirse una mención expresa a la intervención humana -e

[883] Hoofnagle, van der Sloot, y Borgesius, «The European Union general data protection regulation: what it is and what it means», 68.

incluso al término "significativo"- como componente esencial de la toma de decisiones no basada únicamente en el tratamiento automatizado, puesto que es la medida organizativa que legitima determinado tratamiento prohibido con carácter general; también en el apartado primero la cláusula o trabalenguas "que produzca efectos jurídicos en él o le afecte significativamente de modo similar" debería simplificarse, al margen de posibles soluciones -más abajo mencionadas- para interpretar con seguridad jurídica el umbral del riesgo establecido por esta disposición; debería haberse evitado la confusión -y los consiguientes esfuerzos doctrinales- ante la no inclusión del derecho a una explicación -mencionado en el considerando 71- entre las medidas de salvaguarda del apartado tercero.

El origen del sobrenombre kafkiano del artículo 22 RGPD proviene del relato "El proceso", dada la oscuridad que predomina en la toma de decisiones automatizada basada en la elaboración de perfiles. Ahora, hoy en día lo kafkiano de la disposición parece más bien responder a su carácter trágicamente absurdo e ininteligible. Un carácter del que debería desprenderse cualquier norma jurídica que pretenda ser aplicada de forma efectiva.

TERCERA.

Coincido plenamente con Bayamlioglu en que el destino fatal -o no- de las disposiciones del RGPD sobre la elaboración de perfiles y toma de decisiones automatizada depende en gran medida de la pronta reacción del Comité Europeo de Protección de Datos y otras autoridades de la UE[884].

En esta investigación son varias las ocasiones en las que se ha aludido a la oportunidad de que el CEPD realice aportaciones de *soft law* en el marco de sus competencias, por ejemplo: (a) aportando seguridad jurídica para la gestión de riesgos de los responsables del tratamiento a la hora de determinar qué criterios pueden utilizarse para evaluar si un sistema automatizado de elaboración de perfiles produce efectos jurídicos o de afectación significativa similar; o (b) definiendo con mayor detalle el alcance de los derechos de información y acceso cuando se hace uso de estos sistemas, incluyendo: a qué sistemas deben aplicarse los artículos 13(2)(f), 14(2)(g) y 15(1)(h) RGPD -además de a los que

[884] Bayamlıoğlu, «The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"», 17.

adoptan decisiones basados únicamente en el tratamiento automatizado-; a qué clase de explicaciones satisfacen las exigencias del principio de transparencia conforme a los derechos de información y acceso del RGPD; o cómo deben ponderarse las posibles colisiones de derechos legítimos de terceros o limitaciones técnicas razonables con el ejercicio de estos derechos.

Si la conclusión anterior apuntaba al legislador europeo como responsable para la clarificación y simplificación de la regulación de la toma de decisiones automatizada basada en la elaboración de perfiles en el RGPD, en esta ocasión es necesario llamar la atención del CEPD, SEPD y de las propias autoridades de control estatales que, en el marco de sus competencias, disponen de mecanismos más ágiles que el legislador para aliviar la incertidumbre que genera la aplicación de esta regulación.

CUARTA.

Una de las limitaciones de la presente investigación es que no se ocupa directamente de las excepciones contenidas en los apartados 22(2) y 22(4) para legitimar la toma de decisiones basada únicamente en el tratamiento de datos personales, dependiendo de si se tratan o no de categorías especiales de datos en la toma de decisiones automatizada. Dada la amplitud de estas excepciones, la estructura del artículo 22 ha sido calificada por la doctrina de 'castillo de naipes' o 'rodaja de queso suizo', dada la facilidad para esquivar la prohibición general de adoptar decisiones basadas únicamente en el tratamiento en base a las mismas. A pesar de esta limitación metodológica, sí creo que resulta oportuno realizar una reflexión al hilo del análisis realizado sobre la intervención humana como mecanismo de gobernanza.

Al analizar los fundamentos de la intervención humana significativa, se ha puesto de relieve que la Comisión tiene una preocupación por mantener la dignidad humana garantizando que los humanos mantengan el papel principal en la “constitución” de sí mismos[885]. Ahora bien, esta preocupación no incluye -por causa de las amplias excepciones previstas- una prohibición del tratamiento automatizado en sí, tanto si sirve como apoyo a la toma de decisiones o es su fundamento único. Es decir, el RGPD no configura un espacio inviolable y no-computable de la personalidad frente a la

[885] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 84.

automatización, como sí lo hacen otras normas, entre ellas, la propuesta de Reglamento AIA con la prohibición de sistemas de IA: *que se sirva(n) de técnicas subliminales que trasciendan la conciencia de una persona para alterar de manera sustancial su comportamiento de un modo que provoque o sea probable que provoque perjuicios físicos o psicológicos a esa persona o a otra* -5(1)(a) AIA-. En el RGPD no hay límites a la clase de elaboración de perfiles -sobre categorías especiales de datos o no, con unos u otros fines- que pueda servir de base a la toma de decisiones automatizada -basada o no únicamente en el tratamiento-, siempre que el responsable del tratamiento cumpla con la normativa de protección de datos y demuestre su cumplimiento.

Esta conclusión lleva a la cuestión de si, ante la inexistencia de un espacio inviolable y no-computable de la personalidad frente a la automatización o al perfilado, existe o no un riesgo de destrucción de la personalidad en el RGPD en los términos expuestos por Turégano: *la destrucción de la personalidad y el acaparamiento de lo íntimo por la agregación de datos y la generación de perfiles que simplifican en exceso la complejidad de la subjetividad y que no respetan la forma en que el propio sujeto percibe su singularidad, poniéndolo a disposición de intereses que le trascienden*[886].

QUINTA.

Al hilo de la conclusión anterior, y sobre el potencial riesgo de destrucción de la personalidad sobre la "libre" elaboración de perfiles, son varias las barreras normativas fundamentales que el RGPD establece para paliar los efectos de no configurar dicho espacio inviolable y no-computable de la personalidad frente a la automatización.

La primera de esas barreras sería que, por supuesto, un perfilado que vulnere directamente los principios del RGPD, por resultar ilícito o discriminatorio -no leal-, no estaría amparado por la normativa para constituir la base de decisiones automatizadas. No obstante, estos principios como barrera dejan un amplio margen de grises sobre la clase de perfilados que sí podrían constituir dichas decisiones. A este respecto, podríamos volver sobre el análisis realizado sobre las limitaciones del principio de exactitud en relación con el derecho de rectificación, aunque más concretamente recomendaría acudir

[886] Turégano Mansilla, «Los valores detrás de la privacidad», 287.

a las conclusiones del trabajo de Wachter y Mittelstadt sobre el derecho a unas inferencias justas en el RGPD[887].

La segunda de esas barreras sería el consentimiento de la persona interesada. Fundamentalmente porque si las interesadas no consienten, el establecimiento de procesos de toma de decisiones basados únicamente en el tratamiento automatizado se limita de forma considerable. La persona interesada puede negarse a ser objeto de esta clase de decisiones en gran medida no otorgando su consentimiento. Habiendo ya hecho referencia a las limitaciones metodológicas de esta investigación por no abordar directamente el análisis de estas excepciones, no corresponde realizar aquí valoraciones sobre este aspecto. Ahora bien, no creo que a nadie sorprenda que señale la problemática del consentimiento como base legitimadora de esta clase de tratamiento, que han descrito muy bien, entre otros, Jones y Edenberg[888].

La última de las barreras está más directamente relacionada con el objeto de estudio de esta investigación: el responsable del tratamiento puede esquivar la prohibición general de adoptar estas decisiones introduciendo una intervención humana significativa *in the loop*. También, en este sentido, esquivar la segunda barrera -el consentimiento- aquí descrita. A este respecto mi posición en estas conclusiones es clara: la intervención humana no puede funcionar como "solución" de métodos de elaboración de perfiles que son de por sí abusivos. Me remito a las limitaciones de la intervención humana como mecanismo de gobernanza desarrolladas y a la conclusión de que ésta debe configurarse como una pieza más en la cadena organizativa para conseguir un tratamiento automatizado de los datos personales lícito, leal, exacto y responsable, entre otros, exigido por el RGPD. La gobernanza de esta clase de toma de decisiones no puede hacerse depender exclusivamente de la intervención humana.

SEXTA.

Esta investigación propone interpretar la regulación de la toma de decisiones automatizada basada en la elaboración de perfiles sobre tres pilares: la responsabilidad sobre el tratamiento [responsabilidad], la capacidad de la persona interesada para influir

[887] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI».

[888] Vid. Jones y Edenberg, «Troubleshooting AI and Consent».

sobre el tratamiento [permeabilidad] y la posibilidad de observar por parte de terceros que dicho tratamiento es permeable y responsable [transparencia]. Bajo esta estructura se ha puesto de manifiesto que los remedios normativos previstos varían de forma considerable en función de si la toma de decisiones está basada únicamente, o no, en el tratamiento automatizado. Lo cual repercute en la relevancia que adquiere la intervención humana como componente esencial de la toma de decisiones -22(1)- para la gobernanza de la elaboración de perfiles utilizada como un apoyo a la toma de decisiones.

Sin ánimo de desmerecer la importancia de garantizar una intervención humana significativa previa a la producción de efectos jurídicos o significativos para la persona interesada [responsabilidad], no parecen justificadas las diferencias establecidas para los derechos de información [transparencia] y de impugnación [permeabilidad] para esta clase de toma de decisiones. A mi juicio, es pertinente la ampliación de los derechos de información y acceso sobre la lógica algorítmica aplicada y sus consecuencias -13(2)(f), 14(2)(g) y 15(1)(h)-, así como del derecho a impugnar la decisión -22(3)-, como medidas de salvaguarda para las decisiones no basadas únicamente en el tratamiento automatizado que producen efectos jurídicos significativos. Con independencia de que deba garantizarse igualmente una intervención humana previa y significativa para estos sistemas automatizados de apoyo a la toma de decisiones, debe reforzarse el control de las personas interesadas sobre las inferencias que sirven de fundamento a dichos procesos y su transparencia.

SÉPTIMA.

Del mismo modo, tanto para garantizar el control de los responsables sobre las decisiones que adoptan, como para garantizar el control por parte de las personas interesadas sobre las decisiones que les afectan, es necesario reforzar la responsabilidad, permeabilidad y transparencia de los mecanismos de gobernanza existentes. También, los vasos comunicantes entre estos pilares en la regulación de la toma de decisiones automatizada y más allá de las cuestiones referidas a la intervención humana que se desarrollan en las siguientes conclusiones.

En este sentido, y entre otras potenciales medidas, en esta investigación se ha hecho patente la necesidad de incluir de forma explícita un derecho autónomo a una explicación sobre la decisión adoptada vinculado a la posibilidad de impugnar la decisión basada o

no únicamente en el tratamiento automatizado. La persona interesada debe conocer la adopción de una decisión particular que le produce un efecto jurídico o significativo y cómo se ha adoptado.

Por un lado, la explicación sobre este tipo de decisiones debería informarse en el momento de su adopción, es decir, configurarse como un derecho de información sobre la decisión particular adoptada y no meramente como un derecho de acceso, tal y como parece estar actualmente concebido de forma exclusiva para decisiones basadas únicamente en el tratamiento automatizado -15(1)(h)-. Recordemos que el análisis sobre el alcance de los derechos de información y acceso para las decisiones basadas únicamente en el tratamiento automatizado, parece incluir la obligación a revelar información sobre la lógica realmente empleada en la decisión particular, y no sólo sobre la funcionalidad general del sistema de una toma de decisiones automatizada, aunque no se incluyese un derecho a una explicación como tal en la parte dispositiva del RGPD[889].

Por otro lado, acerca de la explicación requerida, se ha puesto de manifiesto la utilidad -especialmente a efectos de rebatir o contestar las decisiones automatizadas- de la clase de justificaciones requeridas por la EIPD, frente a las explicaciones sobre la lógica automatizada del algoritmo que requieren los actuales derechos de información y acceso[890]. La aplicación de esta clase de justificaciones -introducidas por el principio de responsabilidad- a la provisión de explicaciones sobre decisiones particulares, requeriría por parte del responsable de una demostración de que la decisión cumple y respeta el núcleo de la protección de datos y sus principios en el contexto particular que se haya adoptado, más allá de la explicación técnica sobre la lógica de los modelos algorítmicos utilizados. A partir de esta clase de justificaciones se vincularía la responsabilidad sobre el tratamiento de datos personales con la transparencia y, en última instancia, con la permeabilidad del mismo.

OCTAVA.

[889] Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 256; Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 114; Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 93.

[890] Siguiendo líneas de trabajo como las de Henin y Le Métayer, «A framework to contest and justify algorithmic decisions»; Kaminski y Malgieri, «Algorithmic impact assessments under the GDPR: producing multi-layered explanations».

La intervención humana como mecanismo de gobernanza de la toma de decisiones automatizada basada en la elaboración de perfiles ha sido objeto de legítimas críticas que no pueden obviarse si la supervisión humana pretende establecerse como uno de los principios fundamentales para la regulación de los sistemas de IA. La falta de sistematización en la doctrina y jurisprudencia -de estos mecanismos de gobernanza y de los distintos objetivos normativos a los que éstos pueden servir- , así como su deficiente técnica legislativa, sin lugar a duda contribuyen a que dicha intervención sea más vulnerable a esta clase de críticas.

En este contexto, y tal y como se ha ido argumentando en este estudio, debe considerarse urgente y prioritario, por un lado, mejorar la técnica legislativa referida a la inclusión de esta clase de mecanismos de gobernanza. El planteamiento de las siguientes preguntas -sin ánimo exhaustivo- puede resultar útil a estos efectos: ¿cuál es el fundamento bajo el que se incluye esta intervención?; ¿a qué objetivos normativos responde su inclusión?; ¿qué clase de intervención/participación/revisión humana se exige?; ¿cuál es la cualificación exigida -adecuada/efectiva/significativa- y qué criterios determinan dicha cualificación?; ¿para qué sistemas debe ser obligatoria una u otra clase de intervención?; ¿resulta factible -basada en la evidencia- la clase de intervención exigida?; ¿qué coste económico/organizativo tiene la aplicación efectiva de este tipo de intervención y qué impacto puede tener este coste en la efectiva aplicación normativa?; ¿qué otros mecanismos de gobernanza contribuyen al cumplimiento de los objetivos normativos establecidos para la intervención humana?; ¿cómo se interrelacionan entre sí?

Por otro lado, es responsabilidad del resto de agentes jurídicos, en primera instancia, no obviar la existencia de este tipo de mecanismos de gobernanza. Partimos de que, a pesar de la claridad de las Directrices del GT29 en este aspecto, gran parte de la doctrina ha olvidado analizar la prohibición de la toma de decisiones basada únicamente en el tratamiento automatizado del RGPD -22(1)- en clave de mecanismo de gobernanza basado en la intervención humana. Y en segunda instancia, estos agentes jurídicos deben tratar de sistematizar esta clase de mecanismos y de aportar soluciones interpretativas que, ya no los tribunales, sino la sociedad en su conjunto pueda aplicar -pensemos en un responsable del tratamiento que quiera implementar un sistema automatizado basado en la elaboración de perfiles-.

Una vez más, si pretende tomarse en serio la importancia que se prodiga de la supervisión humana como uno de los principios fundamentales para la regulación de los sistemas de IA, estas labores son urgentes y prioritarias, y bajo dicha premisa se ha desarrollado esta investigación.

NOVENA.

La intervención humana en el RGPD se constituye como ineludible para la toma de decisiones automatizada que produce efectos jurídicos o significativos. Tanto si es adoptada como componente esencial de la toma de decisiones -22(1)- que el responsable del tratamiento debe introducir (*in the loop*) para evitar la prohibición general de forma previa a la producción de efectos para el interesado. Como si se adopta como medida de salvaguarda -22(3)- posterior (*out of the loop*) a la toma de decisiones basada únicamente en el tratamiento automatizado. El análisis jurídico realizado sobre qué debe entenderse por intervención humana arroja, en primer lugar, la necesidad de superar un concepto formal de intervención humana debiendo incorporar el criterio cualitativo introducido por las directrices del GT29 y refrendadas por el CEPD: la intervención humana *significativa*. A pesar de las dificultades para delimitar este concepto, el análisis realizado arrojó algunas de las características que forman parte de esta clase de intervención en el RGPD.

Por un lado, las mencionadas directrices definen que la intervención humana significativa debe [a] llevarse a cabo por persona autorizada y competente para modificar la decisión, [b] realizarse sobre un análisis que tenga en cuenta todos los datos disponibles y [c] no conllevar aplicación rutinaria de los resultados algorítmicos. Mientras que en el caso de la intervención humana como medida de salvaguarda debe, además, [d] tener en cuenta la información facilitada por el interesado; lo cual revela el vínculo instrumental existente -22(3)- entre esta clase de intervención humana y el derecho del interesado a expresar su punto de vista respecto del derecho a impugnar la decisión basada únicamente en el tratamiento automatizado[891].

Por otro lado, el análisis teleológico apoyado también en fuentes doctrinales de referencia apunta en varias direcciones complementarias. En un sentido, a que la intervención

[891] Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations», 22; Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems», 1.

humana como componente esencial de la toma de decisiones automatizada -22(1)- encuentra su fundamento en el principio de responsabilidad del RGPD, impidiendo la dejación de responsabilidad del responsable del tratamiento sobre las decisiones que adopta y afectan significativamente a las personas interesadas. De esta forma, podemos vincular la intervención humana con el cumplimiento de los principios del RGPD y su demostración, en este contexto particular, con la adopción de decisiones automatizadas lícitas, leales y exactas -5(1)(a) y (d)-. En otro sentido, la intervención humana se erige como remedio regulatorio al determinismo tecnológico en protección de la dignidad humana. Esta perspectiva se puede vincular en la significancia de la intervención humana con uno de los criterios ya introducidos por las directrices del GT29, esto es, con la necesidad de evitar una aplicación rutinaria de los resultados algorítmicos. Por último, en el caso de la intervención humana como medida de salvaguarda -22(3)-, la intervención humana como capacidad de explicar, de forma detallada e inteligible, cómo se ha llevado a cabo el tratamiento a las personas interesadas (y que pueda llevar a la impugnación efectiva de la decisión), se fundamenta en el control y participación de las personas interesadas sobre las decisiones que les afectan significativamente cuando se toman a partir del tratamiento de sus datos personales.

DÉCIMA.

Tal y como ponen de manifiesto las mencionadas críticas a la intervención humana, la posibilidad de mitigar los efectos perniciosos de esta clase de decisiones desde un mandato normativo que involucre a los seres humanos en estos procesos decisorios es muy limitada. Este interesante debate nos devuelve a la ineludible dicotomía señalada por Favaretto et al., en virtud de la cual los seres humanos son (somos) tanto la causa de los defectos de estas tecnologías, como los supervisores de su correcto funcionamiento[892].

Las conclusiones jurídicas extraídas del análisis de este debate parten de algunas premisas que creo necesario recalcar. Primero, que debe superarse una visión simplista, de poca o ninguna utilidad práctica, muy habitual en este debate que aboga por una ficticia lucha *juicio humano vs. juicio de la máquina*. Segundo, que la intervención humana en fase de implementación de un sistema puede provocar efectos no deseados o perniciosos, pero igualmente -y de forma inevitable- estos sistemas automatizados están también afectados

[892] Favaretto, De Clercq, y Elger, «Big Data and discrimination: perils, promises and solutions. A systematic review», 21.

por la intervención humana de quienes codifican, desarrollan o deciden implementar los mismos. Tercero que, precisamente, una intervención humana defectuosa en fases previas provocará el desarrollo de sistemas defectuosos -por uso de datos de entrenamiento inexactos o insuficientes, metodologías cuestionables o aplicación de métodos de validación deficientes para el uso previsto, entre otros-, cuyos efectos no pueden ser en ningún caso mitigados por una intervención humana posterior. Y cuarto, que a la hora de evaluar si un proceso decisorio es mejor con una determinada clase de intervención humana, con otro tipo de intervención o sin ella, no podemos considerar neutra la definición de lo que es "mejor" o "peor". Y que, además, en todo contexto normativo complejo, encontraremos distintos valores -exactitud, licitud, lealtad, entre otros en el RGPD- compitiendo entre sí a la hora de determinar qué proceso decisorio resulta más satisfactorio.

A partir de estas premisas puede concluirse, en primer lugar, que la intervención humana por sí misma no es suficiente para conseguir una adecuada supervisión -en sentido amplio- sobre los sistemas automatizados: la intervención humana debe configurarse como una pieza más en la cadena organizativa para la gobernanza de sistemas de toma de decisiones automatizados basados en la elaboración de perfiles. En segundo lugar, que la única forma de discernir entre procesos decisorios con intervención humana satisfactoria -en el sentido de ser útil a los objetivos normativos que son establecidos para dicha intervención- es a partir de su validación y demostración efectiva: es decir, resulta necesario acompañar los mecanismos de gobernanza basados en la intervención humana de exigencias normativas relativas a la validación y demostración de la efectividad de dicha intervención.

UNDÉCIMA.

La gestión de riesgos basada en el principio nuclear de la responsabilidad en el RGPD requiere analizar la intervención humana no solo como un mecanismo de gobernanza contenido en la parte dispositiva referida a los derechos de la persona interesada, sino también como una medida organizativa exigida por dicha gestión integral de los riesgos para la toma de decisiones automatizada. Tanto si se trata de decisiones basadas únicamente en el tratamiento automatizado -22(3): *human out of the loop*-, como si no -22(1): *human in the loop*-, el responsable debe justificar cómo se produce la intervención humana en el proceso y cómo ésta resulta significativa y efectiva. Desde esta perspectiva

de control del riesgo, el responsable del tratamiento puede incluso decidir que participe un agente humano en la toma de decisiones, *in the loop*, como medida/garantía organizativa "extra" para la gestión del riesgo en el tratamiento -incluso cuando el tratamiento cumpliese con alguna de las excepciones del artículo 22 para realizar dicho proceso de forma totalmente automatizada-.

Cuando este proceso decisorio implique un alto riesgo para los derechos y libertades de las personas, esta justificación sobre la introducción de la intervención humana como medida organizativa debe incluirse en la realización de la EIPD. Esta investigación muestra que la EIPD -concebida como un instrumento normativo de autoevaluación[893] cuyo valor reside fundamentalmente en conducir a la construcción de mejores sistemas de tratamiento de datos en su conjunto[894]- resulta útil para que el responsable del tratamiento evalúe, y reevalúe si fuera necesario -dado su carácter continuo-, el rol de la intervención humana en el proceso decisorio diseñado, la eficacia de dicha intervención como medida organizativa y, en consecuencia, demuestre el cumplimiento normativo. En este sentido, es destacable el amplio margen de discrecionalidad permitido por el RGPD para el diseño de sistemas de toma de decisiones automatizados basados en la elaboración de perfiles con intervención humana significativa como componente esencial de la toma de decisiones -22(1)-, que permite la búsqueda del mejor modelo de colaboración humano-máquina posible dentro dicho margen.

No obstante, el RGPD es vago a la hora de establecer la metodología a llevar a cabo al realizar la EIPD y tampoco el CEPD -ni previamente el GT29- ha explorado esta intersección entre la intervención humana como mecanismo de gobernanza y como medida organizativa para el diseño de procesos de toma de decisiones automatizada basada en la elaboración de perfiles. Esto último contribuiría a una mayor seguridad jurídica para los responsables del tratamiento en el diseño de dichos procesos.

DUODÉCIMA.

[893] Mantelero, «AI and Big Data: A blueprint for a human rights, social and ethical impact assessment», 768; Hawath, «Regulating Automated Decision-Making: An Analysis of Control over Processing and Additional Safeguards in Article 22 of the GDPR.», 171.

[894] Edwards y Veale, «Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?», 51.

La introducción de la responsabilidad sobre el cumplimiento normativo y su demostración como principio nuclear del RGPD ha consolidado el control colectivo sobre la justificación de los procesos que recopilan, procesan y usan los datos personales. En esta investigación se ha resaltado el valor de este control colectivo para los procesos de toma de decisiones automatizada basados en la elaboración de perfiles, frente a modelos normativos de control individual que desplazan la responsabilidad a personas de por sí atomizadas por estos procesos. No obstante, el sistema normativo de responsabilidad del RGPD puede resultar insuficiente, especialmente cuando el control colectivo se aleja de mecanismos de control directos y abiertos a la sociedad civil.

Coincido con Vedder en que sería deseable un régimen regulatorio que permita deliberar de forma activa sobre los posibles impactos en los seres humanos del tratamiento de datos personales[895]. Sin embargo, en el RGPD las escasas posibilidades de deliberación por la sociedad civil respecto de los procesos de demostración del cumplimiento -EIPD- se hacen depender de la voluntad del propio responsable del tratamiento -35(9)-, que no tiene la obligación de hacer pública la EIPD. De esta forma, bajo el principio de responsabilidad, el proceso de demostración del cumplimiento al que obliga el RGPD puede ser controlado de forma indirecta por el foro -autoridades competentes y judiciales-, pero no prevé mecanismos de control directo sobre el mismo. Bajo mi punto de vista, este aspecto requiere de una revisión normativa.

Resulta necesario reforzar la publicidad [transparencia] de la evaluación de impacto de protección de datos [responsabilidad] para favorecer la deliberación colectiva [permeabilidad] sobre la toma de decisiones automatizada basada en la elaboración de perfiles. Con ello no quiero decir que dicha publicidad deba resultar preceptiva para todo tipo de tratamiento de datos que requiera de la realización de una EIPD. Por ejemplo, sería posible establecer un ulterior umbral normativo sobre la clase de tratamiento que requiere de este tipo de publicidad. Tampoco que la publicidad incluya la totalidad del contenido exigido para la realización de la EIPD, sino únicamente los aspectos que resulten determinantes para comprender la justificación del tratamiento de datos en el contexto particular en el que se implementan los procesos de toma de decisiones automatizados.

[895] Vedder, «Why data protection and transparency are not enough when facing social problems of machine learning in a big data context», 44.

DECIMOTERCERA.

En esta investigación la normativa europea de protección de datos ha sido analizada como referencia para la regulación de la toma de decisiones automatizada basada en la elaboración de perfiles. Más allá de sistemas de toma de decisiones que puedan escapar del ámbito de aplicación de esta normativa y que no han sido objeto de esta investigación -por ejemplo, sistemas que no incluyen el tratamiento de datos personales o aquéllos que se rigen expresamente por la Directiva (UE) 2016/680 en materia penal-, es posible que la aplicación de esta normativa sea insuficiente para la protección de los derechos y libertades de las personas en determinados ámbitos. Toda mejora en la normativa de protección de datos ha de ser bienvenida, no obstante, sería impensable exigir de esta normativa soluciones aplicables a toda clase de sistema de toma de decisiones automatizado y durante todo su ciclo de vida.

Por ello, Janssen destaca que las normativas sectoriales son necesarias para la protección de los derechos fundamentales[896]. Es necesario profundizar en el análisis de las intersecciones normativas existentes a nivel sectorial entre la protección de datos y la regulación de los modelos algorítmicos en todas sus fases; es decir, tanto en el diseño y desarrollo de los mismos -pongamos la normativa sobre productos sanitarios que regula los sistemas que tengan un fin médico específico-, como en su implementación -siguiendo el ejemplo anterior, la normativa sobre autonomía de los pacientes en el entorno clínico-.

Desde luego, esta cuestión es vital en lo que se refiere a las fases de diseño y desarrollo de los modelos, dado que la regulación de la toma de decisiones automatizada en el RGPD no abarca dicha fase. Y puede resultar fundamental también a la hora de demostrar -o no- el cumplimiento normativo de la protección de datos en la implementación posterior del modelo -una vez más, siguiendo con el ejemplo anterior que vislumbra la necesidad de este análisis sectorial, podría plantearse si es lícito, desde la óptica del RGPD, el tratamiento de datos de un sistema con un fin médico que no haya sido certificado conforme a la normativa de productos sanitarios-. Por último, sería interesante poder explorar formas de comunicación entre las distintas autoridades de control que la

[896] Janssen, «An approach for a fundamental rights impact assessment to automated decision-making», 106.

normativa sectorial y la normativa de protección de datos pueda establecer para la rendición de cuentas.

Por supuesto en estas intersecciones serán muy relevantes las novedades normativas propuestas por la Comisión Europea que se refieren a la gobernanza de datos y sistemas de IA. No obstante, sí me gustaría recalcar que la producción normativa de la Comisión en estos ámbitos debe guardar una especial cautela en su interrelación con la normativa vigente. Un "exceso" normativo podría dar lugar a mayor complejidad e incluso duplicidades, produciendo un efecto contraproducente que complique aún más la aplicación normativa respecto de estos sistemas.

DECIMOCUARTA.

Por último, debo referirme a la propuesta de Reglamento AIA de la Comisión Europea. En el análisis jurídico aquí desarrollado se han destacado los aspectos más prometedores de esta propuesta en relación con el objeto de estudio y, en particular, para reforzar la garantía de una intervención humana significativa y demostrable exigida por el RGPD.

Por un lado, esta propuesta centra sus obligaciones en el diseño y desarrollo de los sistemas de IA, al contrario que la regulación de la toma de decisiones automatizada en el RGPD. Lo hace además con un sistema de gestión de riesgos que comparte similitudes con el modelo normativo basado en la responsabilidad del RGPD. Por otro lado, establece la obligación de diseñar y desarrollar los sistemas de alto riesgo de modo que puedan ser supervisados de forma efectiva por personas físicas durante su fase de uso. De esta forma, aunque la propuesta no establece obligación jurídica alguna de supervisión o intervención humana en fase de implementación, las medidas que el proveedor adopte en fase de diseño y desarrollo para asegurar que el sistema puede ser supervisado adecuadamente, serán determinantes para el cumplimiento normativo y su demostración en los mandatos que el Derecho establezca en fases posteriores -por ejemplo, tal y como lo hace el RGPD en su artículo 22-. Es decir, el cumplimiento normativo -obligado por la propuesta de Reglamento AIA- por los proveedores del requisito de supervisión humana en fase de desarrollo contribuirá, a su vez, al referido cumplimiento normativo y su demostración por parte del usuario/responsable del tratamiento que implante el sistema de IA; en este caso, derivado de los mandatos del artículo 22 y 35 RGPD sobre intervención humana para procesos de toma de decisiones automatizada.

No obstante, esta propuesta corre el riesgo de calcificar algunos de los defectos ya existentes en el modelo normativo de responsabilidad adoptado por el RGPD. A mi juicio, este riesgo no se produce tanto en el sentido señalado por la doctrina e instituciones como el CEPD y SEPD[897]; es decir, no tanto por la falta de inclusión de derechos individuales en esta propuesta que permitan a las personas afectadas tener acceso directo [transparencia] y accionable [permeabilidad] al diseño y desarrollo de sistemas que les afectan de forma particular. A mi entender, esta clase de derechos tienen mejor acomodo en la normativa que regula la implementación de esta clase de sistemas -entre otras, en la normativa de protección de datos-.

Sí coincido, no obstante, en el diagnóstico: el modelo de responsabilidad adoptado es ciego en tanto que los requerimientos de transparencia y permeabilidad no son accesibles para la sociedad civil (foro) en su conjunto, al igual que ocurre con la EIPD en el modelo de responsabilidad del RGPD. Y no voy sino a reiterar mi argumento arriba desarrollado. No es necesario que la transparencia y permeabilidad del proceso de gestión de riesgos y documentación incluya el acceso o publicidad a la totalidad de cada proceso de certificación de sistemas de IA de alto riesgo, sino únicamente a los aspectos que resulten determinantes para comprender el diseño y desarrollo de los sistemas de IA en relación con el contexto particular en el que van a ser implementados para la toma de decisiones automatizada basada en la elaboración de perfiles.

CONCLUSIONS ON THE GOVERNANCE AND HUMAN OVERSIGHT OF AUTOMATED DECISION MAKING BASED ON PROFILING

English version

I would not want the theoretical framework developed in this research to offer a reductionist picture of the problems of automated decision-making. According to Hildebrandt's claims, reducing the issue of implementing algorithmic models to their "black box" or "opaque" attributes, to the potential biases they might introduce or

[897] Veale y Zuiderveen Borgesius, «Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach»; Cotino et al., «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)»; Comité Europeo de Protección de Datos (CEPD) y Supervisor Europeo de Protección de Datos (SEPD), «Dictamen conjunto 5/2021 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)».

reinforce, or to the potential implementation of human oversight over them, may distract attention from the underlying and fundamental problems with these processes[898]. Mainly, the trade-offs inherent in the introduction of algorithmic logics in different decision-making processes in our social environment; automating, datifying -quantifying- and regulating under statistical parameters, to a greater or lesser extent, the human activity in which it is decided to implement a system of these characteristics.

In any case, I do hope that the theoretical framework developed will be helpful to scholars who wish to approach the topic as a tool for broadly understanding what stages comprise the decision-making systems: what phases integrate automated decision-making systems based on profiling and their legal relevance; the different ways in which the law can involve humans in these processes to serve different policy objectives; what biases are, how they are reproduced in an automated system and their different statistical, ethical and legal significance; the different kinds of opacity that can be referred to when talking about the black box effect of algorithmic models, and how each kind of opacity calls for a different legal response appropriate to the context in which a system is implemented.

This opening comment on the first chapter of this work gives way then to the final conclusions drawn from the analysis of European data protection law for the governance and human oversight of automated decision-making based on profiling.

FIRST. Looking at the regulation of automated decision-making based on profiling in the GDPR from a theoretical perspective, it is possible to conclude that, while many governments and technology companies wish to eliminate humans from decision-making processes, the GDPR goes in the opposite direction and shows in Article 22 a clear policy preference for placing humans in the loop[899]. In this regard, the data protection regulation is fully aligned with the EU's regulatory initiatives on high-risk AI systems, which have without exception placed human oversight as a core and mandatory requirement for the development and implementation of these systems. However, from a more down-to-earth perspective on the real-world influence of the GDPR, a further failure of European data protection law to ensure compliance with the regulation of automated decision-making emerges, as this work demonstrates. This failure is even more obvious when it comes to

[898] Hildebrandt, «Data-Driven Prediction of Judgment. Law's New Mode of Existence?»

[899] Hoofnagle, van der Sloot, y Borgesius, «The European Union general data protection regulation: what it is and what it means», 68.

governance mechanisms based on human intervention. Concerning these mechanisms, with little relevance in the already limited case law interpreting Article 22, the literature has focused the analysis on human intervention as a safeguard measure for decision-making based solely on automated processing -22(3): human out of the loop-, rather than on human intervention as an essential component for decision-making not based solely on automated processing -22(1): human in the loop-.

SECOND.

It is indisputable that the Article 22's wording lacks the clarity and simplicity required for any legal rule that is intended to be effectively applied, aside from other issues that may also contribute to the failure of this provision. Not to be exhaustive, as the improvements here could be very different depending on the kind of solution to be found; the first and fourth paragraphs should specify that these are two general prohibitions for the controller on taking decisions based solely on automated processing -and not a right to be exercised on request by the data subject-; the first paragraph should include an explicit reference to human intervention -and even the term 'meaningful'- as an essential component of decision-making not based solely on automated processing, given that it is the organisational measure that justifies a generally prohibited processing operation; in the first paragraph too, the clause or tongue twister " which produces legal effects concerning him or her or similarly significantly affects him or her" should be simplified, regardless of the possible alternative solutions -see below- to interpret with legal certainty the risk threshold established by this provision; the misunderstanding on the right to an explanation -mentioned in Recital 71 but not included among the safeguard measures in the third paragraph- should be avoided.

The origin of the Kafkaesque alias for Article 22 GDPR comes from "The Process", given the obscurity that prevails in automated decision-making based on profiling. However, today, the Kafkaesque nature of the provision seems rather to reflect its tragically absurd and unintelligible character. Such character is unbearable for any legal rule that seeks to be effectively enforced.

THIRD.

I wholly agree with Bayamlioglu that the quick response of the European Data Protection Board and other EU authorities will play a significant role in whether or not the GDPR's rules on profiling and automated decision-making meet their cursed fate[900].

This research has on several occasions alluded to the opportunity for the EDPB to provide soft law contributions within its powers, e.g. by: (a) providing legal certainty for risk management by data controllers when determining which criteria may be used to assess whether an automated profiling system produces similar legal effects or significant impact; or defining in more detail the scope of information and access rights when such systems are used, including (b) to which systems 13(2)(f), 14(2)(g) and 15(1)(h) GDPR should apply -in addition to those that make decisions based solely on automated processing-; (c) what kind of explanations satisfy the requirements of the transparency principle under the information and access rights of the GDPR; or (d) how possible collisions of legitimate rights of third parties or reasonable technical limitations with the exercise of these rights should be weighed against the exercise of these rights.

While the previous conclusion pointed to the European legislator as responsible for the clarification and simplification of the regulation of automated decision-making based on profiling in the GDPR, on this occasion it is necessary to draw the attention of the EDPB, EDPS and the national data protection authorities themselves who, within the scope of their competences, have more agile mechanisms than the legislator to alleviate the uncertainty generated by the application of this regulation.

FOURTH.

One of the limitations of this research is that it does not directly address the exceptions contained in paragraphs 22(2) and 22(4) to legitimise decision-making based solely on the processing of personal data, depending on whether special categories of personal data are processed or not in automated decision-making. Given the breadth of these exceptions, the structure of Article 22 has been described by legal scholars as a 'castle of cards' or a 'slice of Swiss cheese'. These exceptions allow for easy circumvention of the general prohibition on making decisions based solely on processing. Despite this

[900] Bayamlıoğlu, «The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called “right to explanation”», 17.

methodological limitation, I do think it is worth reflecting on this issue in the context of the analysis of human intervention as a governance mechanism.

Looking at the rationale for meaningful human intervention, it becomes clear that the Commission is concerned with maintaining human dignity by ensuring that humans retain the primary role in "constituting" themselves[901]. Nevertheless, this concern does not include -because of the broad exceptions provided- a prohibition of automated processing per se, whether it supports decision-making or is the sole basis for it. In other words, the GDPR does not create an inviolable and non-computable space of personality from automation, as do other rules, like the proposed AIA Regulation with the prohibition of any AI system: *that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm* -5(1)(a) AIA-. There are no limits in the GDPR on the kind of profiling -based on special categories of data or not, for one purpose or another- that can be used as a basis for automated decision-making -whether or not based solely on processing- as long as the controller complies with the regulation and demonstrates compliance.

This conclusion leads me to wonder whether in the absence of an inviolable and non-computable space of the personality in the face of automation or profiling, there is a risk of destruction of the personality in the GDPR in the terms set out by Turégano: *the destruction of personality and the monopolisation of the intimate by the aggregation of data and the generation of profiles that oversimplify the complexity of subjectivity and that do not respect the way in which the individuals themselves perceive their uniqueness, placing them at the disposal of interests that transcend them*[902].

FIFTH.

In light of the previous conclusion, and regarding the potential risk of destruction of the personality of "free" profiling, there are several fundamental regulatory barriers that the GDPR establishes to mitigate the effects of not configuring an inviolable and non-computable space of the personality in the face of automation.

[901] Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 84.

[902] Turégano Mansilla, «Los valores detrás de la privacidad», 287.

The first of these barriers would be that, of course, profiling that directly violates the principles of the GDPR, because it is unlawful or discriminatory - not fair - would not be covered by the regulation to form the basis for automated decisions. However, these principles as a barrier leave a wide margin of grey on the kind of profiling that could constitute such decisions. In this respect, we could return to the analysis of the limitations of the accuracy principle in relation to the right of rectification, although more specifically I would recommend referring to the conclusions of Wachter and Mittelstadt's work on the right to reasonable inferences in the GDPR[903].

The second of these barriers would be the data subject's consent. Primarily because if data subjects do not consent -opt in-, the establishment of decision-making processes based solely on automated processing is considerably limited. The data subject may refuse to be subject to such decisions to a large extent by not giving consent. Having already referred to the methodological limitations of this research by not directly addressing the analysis of these exceptions, it is not appropriate here to make assessments on this aspect. However, I do believe that it will not come as a surprise to anyone that I point to the problem of consent as a legitimising basis for this kind of data processing, which has been well described, among others, by Jones and Edenberg[904].

The last of the barriers more directly relates to the scope of this research: the controller can circumvent the general prohibition on making such decisions by introducing a significant human intervention in the loop. And it -a human in the loop- allows to circumvent the second barrier -consent- too. On this point, my position on these findings remains quite clear: human intervention cannot function as a remedy for profiling methods that are in themselves abusive. Here I would recall the limitations of human intervention as a governance mechanism developed and the conclusion that human intervention must be configured as just another more piece in the organisational chain to achieve lawful, fair, accurate and accountable automated processing of personal data, as required by the GDPR. The governance of this kind of decision-making cannot rely solely on human intervention.

[903] Wachter y Mittelstadt, «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI».

[904] Jones y Edenberg, «Troubleshooting AI and Consent».

SIXTH.

This research proposes to interpret the regulation of automated decision-making based on profiling on three pillars: responsibility for the processing [accountability], the ability of the data subject to influence the processing [permeability] and the possibility for third parties to observe that the processing is permeable and accountable [transparency]. Under this scheme, it becomes clear that the regulatory remedies envisaged differ considerably depending on whether the decision-making is based solely or not on automated processing. This results in increased relevance for human intervention as an essential component of decision-making -22(1)- for the governance of profiling used as a decision-support tool.

Without undermining the importance of ensuring meaningful human intervention prior to the production of legal or significant effects for the data subject [accountability], the differences established for the rights of information [transparency] and to contest [permeability] for this kind of decision-making do not seem justified. An extension of the rights of information and access to the algorithmic logic applied and its consequences -13(2)(f), 14(2)(g) and 15(1)(h)- as well as of the right to contest the decision -22(3)- as safeguards for decisions not based solely on automated processing which produce significant legal effects is, in my view, justified and desirable. Irrespective of the fact that prior meaningful human intervention is also to be ensured for such automated decision support systems, the people's control and influence over the inferences underlying such processes and their transparency should be strengthened.

SEVENTH.

In the same way, both to guarantee the control of decision-makers over the decisions they take, and to guarantee control by the data subjects over the decisions that affect them, it is necessary to reinforce the accountability, permeability and transparency of the existing governance mechanisms. Also, the communicating vessels between these pillars in the regulation of automated decision-making and beyond the issues of human intervention which are developed in the following conclusions.

Among other potential measures, this research revealed the need to explicitly include an autonomous right to an explanation of the decision taken, linked to the possibility of contesting the decision whether or not it is based solely on automated processing. The

data subject must be aware of the adoption of a particular decision that has a legal or significant effect on him/her and how it was taken.

On the one hand, the explanation of such decisions should be provided at the time of their adoption, i.e. it should be a right of information about the particular decision taken and not merely a right of access, as it seems to be currently conceived exclusively for decisions based solely on automated processing - 15(1)(h) -. Let me recall that the analysis of the scope of the rights of information and access for decisions based solely on automated processing seems to include the obligation to disclose information about the logic actually employed in the particular decision, and not only about the general system functionality of automated decision-making, even if no right to an explanation as such was included in the provisions of the GDPR[905].

On the other hand, as regards the explanation required, it has become clear that the kind of justifications required by DPIAs are useful -especially for the purpose of challenging or contesting automated decisions- as opposed to the explanations of the automated logic of the algorithm that are required by existing rights of information and access[906]. When applied to the provision of explanations on particular decisions, this kind of justifications -introduced by the accountability principle- would require from the controller a demonstration that the decision complies with and respects the core of data protection and its principles in the particular context it is adopted, beyond the technical explanation on the logic of the algorithmic models used. Based on such justifications, accountability for the processing of personal data would be linked to transparency and, ultimately, to the permeability of the data processing.

EIGHT.

Human intervention as a governance mechanism for automated decision-making based on profiling has been legitimately criticised. These critiques should not be ignored if human oversight is to be established as one of the fundamental principles for the

[905] Malgieri y Comandé, «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation», 256; Brkan, «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond», 114; Mendoza y Bygrave, «The Right Not to be Subject to Automated Decisions Based on Profiling BT -», 93.

[906] Following research approaches such as those of Henin y Le Métayer, «A framework to contest and justify algorithmic decisions»; Kaminski y Malgieri, «Algorithmic impact assessments under the GDPR: producing multi-layered explanations».

regulation of AI systems. The lack of systematisation in the literature and jurisprudence -of these governance mechanisms and of the different regulatory objectives they may serve- as well as their poor legislative technique, undoubtedly contribute to making such intervention more vulnerable to this criticism.

In this context, on the one hand, to improve the legislative technique regarding the inclusion of this kind of governance mechanisms should be considered urgent and a priority. The following non-exhaustive questions may be useful for this purpose: What is the rationale for this intervention's inclusion? What normative goals does its inclusion serve? What type of human involvement or review is necessary? What is the required qualification -adequate/effective/significant- and what criteria determine this qualification? For which systems should one or the other kind of intervention be mandatory? Is the required intervention feasible? Based on evidence? What is the economic/organisational cost of the effective implementation of this kind of intervention and what impact can this cost have on policy implementation? What other governance mechanisms contribute to meeting the policy goals set for human intervention? How do they interrelate with each other?

On the other hand, it is the responsibility of the other legal actors, in the first instance, not to ignore the existence of such governance mechanisms. Despite the clarity of the WG29 Guidelines in this regard, much of the legal scholarship has neglected to analyse the GDPR's 22(1) prohibition of automated processing-only decision-making as a governance mechanism based on human intervention. And secondly, to try to systematise this kind of mechanism and to provide interpretative solutions that, no longer the courts, but society can apply -think of data controllers who want to implement an automated system based on profiling-.

Once again, if the importance of human oversight as one of the fundamental principles for the regulation of AI systems is to be taken seriously, this work is urgent and a priority, and it is under this premise that this research has been conducted.

NINTH.

Human intervention in the GDPR is considered unavoidable for automated decision-making that produces legal or significant effects. Whether it is adopted as an essential component of the decision-making -22(1)- that the controller must introduce (in the loop)

to avoid the general prohibition prior to the production of effects for the data subject. Or it is adopted as a safeguard measure -22(3)- subsequent (out of the loop) to the decision-making based solely on automated processing. The legal analysis carried out on what should be understood by human intervention yields, first, the need to go beyond a formal concept of human intervention by incorporating the qualitative criterion introduced by the WG29 guidelines endorsed by the EDPB: meaningful human intervention. Despite the difficulties in delimiting this concept, the analysis revealed some of the characteristics that constitute this type of intervention in the GDPR.

On the one hand, the aforementioned guidelines establish that meaningful human intervention must [a] be carried out by a person authorised and competent to modify the decision, [b] be based on an analysis that takes into account all available data and [c] not entail routine application of algorithmic results. Whereas in the case of human intervention as a safeguard measure it must, in addition, [d] take into account the information provided by the data subject, which reveals the instrumental link - 22(3) - between this kind of human intervention and the right of the data subject to express his or her point of view, with respect to the right to contest the decision based solely on automated processing[907].

On the other hand, the teleological analysis also supported by referential doctrinal sources points in complementary directions. In one sense, human intervention as an essential component of automated decision-making -22(1)- finds its basis on the accountability principle of the GDPR, preventing the controller from abdicating responsibility for the decisions he or she adopts that significantly affect data subjects. In this respect, we can link human intervention to compliance with the GDPR principles and its demonstration, and for this context, with lawful, fair and accurate automated decision-making -5(1)(a) and (d)-. In another sense, human intervention arises as a regulatory remedy to technological determinism in protection of human dignity. This perspective can be linked to the meaningfulness of human intervention with one of the criteria already introduced by the WG29 guidelines, i.e. the need to avoid a routine application of algorithmic results. Finally, regarding human intervention as a safeguard measure -22(3)- human intervention as the ability to explain, in a detailed and intelligible way, how the processing has been

[907] Malgieri, «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations», 22; Almada, «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems», 1.

carried out to the data subjects (and which can lead to an effective contestation of the decision), is based on the control and involvement of data subjects over decisions that significantly affect them when taken on the basis of the processing of their personal data.

TENTH.

As the aforementioned criticisms of human intervention make clear, it is hardly possible to mitigate the pernicious effects of these kinds of decisions from a normative prescription that involves human beings in these decision-making processes. This interesting debate brings us back to the inescapable dichotomy pointed out by Favaretto et al. whereby humans are (we are) both the cause of the flaws of these technologies and the overseers of their proper functioning[908].

The conclusions drawn from the analysis of this debate are based on several premises that I believe are worth noting. First, the simplistic view, of little or no practical use, which is very common in this debate and which advocates a fictitious competition between human and machine judgement, must be overcome. Second, that human intervention at the implementation stage of a system can cause undesired or pernicious effects, but equally - and inevitably - these automated systems are also affected by the human intervention of those who code, develop or decide to implement them. Thirdly, flawed human intervention in previous phases will lead to the development of flawed systems - due to the use of inaccurate or insufficient training data, questionable methodologies or the application of poor validation methods for the intended use, among others - whose effects can never be mitigated by subsequent human intervention. And fourth, when assessing whether a decision-making process is better with a particular kind of human intervention, with another kind of intervention or without it, we cannot consider the definition of what is "better" or "worse" to be neutral. Furthermore, in any complex regulatory context, we will find different values -accuracy, lawfulness, fairness, among others in the GDPR- competing with each other when determining which decision-making process is more satisfactory.

On these premises, it can be concluded, firstly, that human intervention alone is not sufficient to achieve adequate oversight - in a broad sense - over automated systems:

[908] Favaretto, De Clercq, y Elger, «Big Data and discrimination: perils, promises and solutions. A systematic review», 21.

human intervention must be configured as just another piece in the organisational chain for the governance of automated decision-making systems based on profiling. Secondly, the only way to discern between decision-making processes with satisfactory human intervention - in the sense of being useful to the normative goals that are set for such intervention - is on the basis of its actual validation and demonstration: that is, it is necessary to complement governance mechanisms based on human intervention with normative requirements regarding the validation and demonstration of the effectiveness of such intervention.

ELEVENTH.

Risk management based on the core principle of accountability in the GDPR requires analysing human intervention not only as a governance mechanism contained in the operative part referring to the rights of the data subject but also as an organisational measure required by such holistic risk management for automated decision-making. Whether the decisions are based solely on automated processing -22(3): human out of the loop- or not -22(1): human in the loop- the controller must justify how human intervention in the process takes place and how it is meaningful and effective. Within this risk control perspective, the controller may even decide to involve a human agent in the decision-making process, in the loop, as an 'extra' organizational measure for the processing -even if it meets one of the Article 22 exceptions to make such decisions in a fully automated way-.

Where this decision-making process involves a high risk to the rights and freedoms of individuals, this justification for the introduction of human intervention as an organisational measure should be included when carrying out the DPIA. This research shows that the DPIA -conceived as a normative self-assessment tool[909] whose value mainly lies in leading to the construction of better overall data processing systems[910]- is useful for the controller to assess, and reassess if necessary -given its continuous nature-, the role of human intervention in the designed decision-making process, the effectiveness of such intervention as an organisational measure and, consequently, to

[909] Mantelero, «AI and Big Data: A blueprint for a human rights, social and ethical impact assessment», 768; Hawath, «Regulating Automated Decision-Making: An Analysis of Control over Processing and Additional Safeguards in Article 22 of the GDPR.», 171.

[910] Edwards y Veale, «Enslaving the Algorithm: From a “Right to an Explanation” to a “Right to Better Decisions”?», 51.

demonstrate compliance with the GDPR. In this regard, the wide margin of discretion allowed by the GDPR for the design of automated decision-making systems based on profiling with significant human intervention as an essential component of decision-making -22(1)- is noteworthy, allowing the exploration of the best possible model of human-machine collaboration within that margin.

However, the GDPR is vague in setting out the methodology to be carried out when conducting the EIPD and neither the EDPB -nor previously the WG29- have explored this intersection between human intervention as a governance mechanism and as an organisational measure for the design of automated decision-making processes based on profiling. The latter would contribute to enhanced legal certainty for data controllers in the design of such processes.

TWELFTH.

The introduction of accountability for compliance and its demonstration as a core principle of the GDPR has consolidated collective control over the justification of processes that collect, process and use personal data. This research highlighted the value of this collective control for automated decision-making processes based on profiling, as opposed to regulatory models of individual control that transfer responsibility to individuals who are already atomised by these processes. However, the GDPR's normative accountability scheme may be insufficient, especially when collective control moves away from direct control mechanisms that are open to civil society.

Like Vedder, I agree that a regulatory regime that allows for active deliberation on the potential human impacts of personal data processing would be desirable[911]. However, under the GPDR, the limited possibilities for deliberation by civil society with regard to the processes of demonstrating compliance -DPIAs- are left to the will of the controllers themselves -35(9)- who have no obligation to make the DPIAs public. Thus, under the principle of accountability, the process of demonstrating compliance required by the GDPR can be indirectly controlled by the forum -courts and other judicial authorities- but

[911] Vedder, «Why data protection and transparency are not enough when facing social problems of machine learning in a big data context», 44.

it does not provide for direct control mechanisms over it. In my view, this aspect requires a regulatory review.

It is necessary to strengthen the publicity [transparency] of the data protection impact assessment [accountability] in order to encourage collective deliberation [permeability] on automated decision-making based on profiling. This is not to say that such disclosure should be mandatory for every type of data processing that requires a DPIA to be carried out. For example, it would be possible to set a further regulatory threshold on the kind of processing that requires such disclosure. Nor should the disclosure include the full content required for carrying out the DPIA, but only those aspects that are essential to understand the justification for the data processing in the particular context in which automated decision-making processes are implemented.

THIRTEENTH.

The European data protection law was analysed in this study as the reference for the regulation of automated decision-making based on profiling. Beyond decision-making systems that may fall outside the scope of this regulation and have not been the subject of this research - for example, systems that do not involve the processing of personal data or those that are expressly governed by Directive (EU) 2016/680 on criminal matters - the application of this regulation may be insufficient for the protection of the rights and freedoms of individuals in certain areas. Any improvement in data protection regulation is to be welcomed, however, it seems unreasonable to demand from this regulation that it should provide solutions for all kinds of automated decision-making systems and throughout their lifecycle.

For this reason, Janssen stresses that sector-specific regulations are essential for the protection of fundamental rights[912]. It is necessary to deepen the analysis of the regulatory intersections existing at the sectorial level between data protection and the regulation of algorithmic models in all their stages; that is, both in their design and development -let us take the regulation on medical devices that regulates systems that have a specific medical purpose- and in their implementation -following the previous example, the regulation on patient autonomy in the clinical context-.

[912] Janssen, «An approach for a fundamental rights impact assessment to automated decision-making», 106.

Certainly, this issue is critical in terms of the design and development phases of the models, as the regulation of automated decision-making in the GDPR does not address this stage. And it may also be crucial when it comes to demonstrating -or not- data protection compliance in the subsequent implementation of the model -again, following on from the previous example that hints at the need for this sectorial analysis, one could ask whether it is lawful under the GDPR to process data from a system for a medical purpose that has not been certified as compliant with the medical devices regulation-. Finally, it would be interesting to be able to explore ways of communication between the different supervisory authorities that the sectorial regulation and the data protection regulation establish to ensure accountability.

In these intersections, the new regulatory developments proposed by the European Commission that refer to data governance and AI systems will of course be very relevant. However, I would like to note that the Commission's regulatory output in these areas needs to be particularly cautious in its interplay with existing regulations. An over-regulation could lead to further complexity and even overlaps, producing a counterproductive effect that further complicates the regulatory enforcement with respect to these systems.

FOURTEENTH.

Finally, I must refer to the European Commission's AIA Regulation proposal. The legal analysis developed here highlighted the most promising aspects of this proposal regarding the object of study and, in particular, to reinforce the guarantee of meaningful and demonstrable human intervention required by the GDPR.

On the one hand, this proposal focuses its obligations on the design and development of AI systems, contrary to the regulation of automated decision-making in the GDPR. It does so with a risk management system that shares similarities with the GDPR's accountability-based regulatory model. On the other hand, it establishes the obligation to design and develop high-risk systems in such a way that they can be effectively overseen by natural persons during their usage phase. Thus, although the proposal does not establish any legal obligation for human oversight or intervention at the deployment or usage stage, the measures that the provider adopts at the design and development stage to ensure that the system can be adequately overseen, will become crucial for compliance and its

demonstration under the mandates that the law establishes at later stages -for example, as the GDPR does in Article 22-. In other words, regulatory compliance - mandated by the proposed AIA Regulation - by providers with the human oversight requirement at the development stage will, in turn, contribute to regulatory compliance and its demonstration by the user/processor who implements the AI system; in this case, stemming from the mandates of Articles 22 and 35 GDPR on human intervention for automated decision-making processes.

However, this proposal carries the risk of calcifying some of the flaws already existing in the regulatory model of accountability adopted by the GDPR. In my view, this risk is not so much in the sense pointed out by some legal scholars and institutions such as the EDPB and EDPS[913]; that is, not due to the lack of individual rights included in this proposal that would allow affected individuals to have direct [transparency] and actionable [permeability] access to the design and development of systems that affect them in a particular way. In my opinion, these kinds of rights are better accommodated within the laws that regulate the implementation of these kinds of systems -among others, in data protection law-.

I do agree, however, with their diagnosis: the accountability model adopted is blind as long as the transparency and permeability requirements are not accessible to civil society (forum) in general, as is the case with the DPIAs under the GDPR's accountability model. And I shall do no more than restate the point made above. Transparency and permeability of the risk management and documentation process need not include access or disclosure to the entire certification process of high-risk AI systems, but only to those aspects that are decisive for understanding the design and development of AI systems in relation to the particular context in which they are to be implemented for automated decision-making based on profiling.

[913] Veale y Zuiderveen Borgesius, «Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach»; Cotino et al., «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)»; Comité Europeo de Protección de Datos (CEPD) y Supervisor Europeo de Protección de Datos (SEPD), «Dictamen conjunto 5/2021 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)».

UN EPITAFIO MÁS QUE UN EPÍLOGO

UN EPITAFIO MÁS QUE UN EPÍLOGO

Esto es un poco todo.

Comencé esta etapa en noviembre de dos mil diecisiete. Por puro desconocimiento eché a andar con un concepto sobre la investigación, la ciencia y la universidad muy diferente del que hoy manejo. Ni mal ni bien, digo.

Ahora vendrá la ANECA. Me veo un poco en bragas porque no he realizado contribución alguna en congresos de innovación docente.

Además, imaginaos el estrés por aquel congreso en Namur en el que se me olvidó pedir el certificado de asistencia. Mi nombre está en el programa, conservo los mails de aceptación, mi conferencia grabada en vídeo y mi contribución en el libro colectivo, pero quién sabe si será suficiente. Quizás rescate al efecto alguna foto de la cena de gala tras la cuarta copa de vino.

Tampoco estoy yo para quejarme de burocracia, que aún estoy en nivel principiante.

Por otro lado, la producción científica bate récords imposibles para satisfacción de gestores universitarios que pelean por estar en *rankings* de renombre. Entre tanto, la crisis de reproducibilidad de la ciencia tiene sus ecos en las llamadas ciencias jurídicas. Escribimos más que nunca para que nadie nos lea. Y como sabemos que, a lo sumo, nos leerán de pasada, terminamos escribiendo también de pasada.

No queda un jurista vivo que no haya escrito sobre "IA" en los últimos 5 años. Quién soy yo para juzgar a nadie, espero haberos citado aquí a todos y todas. A los miembros del tribunal, me refiero.

Dejando ya el cinismo a un lado, y parafraseando a Ursula K. Le Guin, vivimos en un sistema de cuyo poder puede parecer imposible escapar. También parecía imposible hacerlo del derecho divino de los monarcas. Sin embargo, cualquier poder humano puede ser resistido y cambiado.

Como de cualquier otra experiencia humana que he transitado, lo mejor que me llevo son las personas que me han acompañado en esta vivencia. No me cabían todas en los agradecimientos, y tampoco lo harían aquí.

Pero son esas personas las que me han concedido una pausa. Una pausa que me ha permitido pensar, aunque sea un ratito, para ser consciente de dónde me hallo. Y a partir de esa consciencia tratar de decidir a dónde quiero llegar con, bueno, una poca honestidad hacia las demás y hacia mí mismo.

Y aquí he llegado. A un epitafio más que a un epílogo.

El de, y aunque sea amor, el amor es una pena también, en el fondo, era Camarón. Terminemos con esto, que queda mucha casa por barrer.

REFERENCIAS BIBLIOGRÁFICAS[914]

[914] En lo que respecta a las fuentes normativas y jurisprudenciales, las notas al pie incluyen toda la información necesaria para su identificación, razón por la cual no se incluyen en este apartado. Tampoco se incluye aquí la bibliografía relativa a las conclusiones debido a que no hay aportación doctrinal alguna que resulte novedosa en relación con los anteriores capítulos.

BIBLIOGRAFÍA PRESENTACIÓN: OBJETO DE ESTUDIO, JUSTIFICACIÓN Y MÉTODO

Agencia de los Derechos Fundamentales de la Unión Europea (FRA). «Getting the future right - Artificial intelligence and fundamental rights». Luxemburgo, 2020. https://doi.org/10.2811/774118.

Araujo, Theo, Natali Helberger, Sanne Kruikemeier, y Claes H de Vreese. «In AI we trust? Perceptions about automated decision-making by artificial intelligence». *AI & SOCIETY* 35, n.º 3 (2020): 611-23. https://doi.org/10.1007/s00146-019-00931-w.

Castelluccia, Claude, y Daniel Le Métayer. «Understanding algorithmic decision-making: Opportunities and challenges». Bruselas, 2019. https://doi.org/10.2861/536131.

Djeffal, Christian. «AI, Democracy and the Law». En *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*, editado por Andreas Sudmann, 255-84. transcript Verlag, 2020. https://doi.org/doi:10.1515/9783839447192-016.

Floridi, Luciano, y Mariarosaria Taddeo. «What is data ethics?» *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, n.º 2083 (2016): 20160360. https://doi.org/10.1098/rsta.2016.0360.

Gómez-González, Emilio, Emilia Gomez, Javier Márquez-Rivas, Manuel Guerrero-Claro, Isabel Fernández-Lizaranzu, María Isabel Relimpio-López, Manuel E. Dorado, María José Mayorga-Buiza, Guillermo Izquierdo-Ayuso, y Luis Capitán-Morales. «Artificial intelligence in medicine and healthcare: a review and classification of current and near-future applications and their ethical and social Impact». Luxemburgo, 2020. https://doi.org/10.2760/047666.

Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (HLEG-AI). *Directrices éticas para una IA fiable*. Editado por Comisión Europea. Bruselas: Oficina de Publicaciones de la Comisión Europea, 2019. https://doi.org/doi/10.2759/14078.

Gutwirth, Serge, y Paul De Hert. «Regulating Profiling in a Democratic Constitutional State». En *Profiling the European Citizen: Cross-Disciplinary Perspectives*, editado por Mireille Hildebrandt y Serge Gutwirth, 271-302. Dordrecht: Springer Netherlands, 2008. https://doi.org/10.1007/978-1-4020-6914-7_14.

Krupiy, Tetyana (Tanya). «A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective». *Computer Law & Security Review* 38 (2020): 105429.

https://doi.org/https://doi.org/10.1016/j.clsr.2020.105429.

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, y Patrick Vinck. «Fair, Transparent, and Accountable Algorithmic Decision-making Processes». *Philosophy & Technology* 31, n.º 4 (2018): 611-27. https://doi.org/10.1007/s13347-017-0279-x.

McQuillan, Dan. «The Political Affinities of AI». En *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*, editado por Andreas Sudmann, 163-74. transcript Verlag, 2020. https://doi.org/doi:10.1515/9783839447192-010.

Nettesheim, Martin, y Benedikt Quarthal. «La reforma de las Constituciones de los Länder». *Revista de Estudios Políticos* 151 (2011): 281-310.

Spielkamp (Ed.), Matthias. «Automating Society. Taking Stock of Automated Decision-Making in the EU», 2019.

Taddeo, Mariarosaria. «Trusting Digital Technologies Correctly». *Minds and Machines* 27, n.º 4 (2017): 565-68. https://doi.org/10.1007/s11023-017-9450-5.

BIBLIOGRAFÍA INTRODUCCIÓN A LA GOBERNANZA Y SUPERVISIÓN HUMANA DE LA TOMA DE DECISIONES AUTOMATIZADA BASADA EN LA ELABORACIÓN DE PERFILES

Agencia de los Derechos Fundamentales de la Unión Europea (FRA). «Getting the future right - Artificial intelligence and fundamental rights». Luxemburgo, 2020. https://doi.org/10.2811/774118.

Allo, Patrick. «Mathematical Values And The Epistemology Of Data Practices». En *BEING PROFILED*, 2019. https://doi.org/10.1515/9789048550180-004.

Ben-Shahar, Omri. «Data Pollution». *Journal of Legal Analysis* 11 (1 de enero de 2019): 104-59. https://doi.org/10.1093/jla/laz005.

Bentley, Peter J., Miles Brundage, Olle Häggström, y Thomas Metzinger. «¿Debemos temer a la inteligencia artificial? Análisis en profundidad». Bruselas, 2018. https://doi.org/10.2861/61195.

Boden, Margaret A. *Inteligencia Artificial*. Madrid: Turner, 2017.

Boix, Andrés. «Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones». *Revista de Derecho Público: Teoría y método* 1 (2020): 223-70. https://doi.org/https://doi.org/10.37417/RPD/vol_1_2020_33.

Bostrom, Nick, y Eliezer Yudkowsky. «The ethics of artificial intelligence». En *The Cambridge Handbook of Artificial Intelligence*, 2014. https://doi.org/10.1017/cbo9781139046855.020.

Boucher, Philip. «Artificial intelligence: How does it work, why does it matter, and what can we do about it?» Bruselas, 2020. https://doi.org/10.2861/44572.

Boyd, Danah, y Kate Crawford. «Critical Questions for Big Data». *Information, Communication & Society* 15, n.º 5 (2012): 662-79. https://doi.org/10.1080/1369118x.2012.678878.

Bringsjord, Selmer, y Naveen Sundar Govindarajulu. «Artificial Intelligence». En *The Stanford Encyclopedia of Philosophy (Winter 2019 Edition)*, editado por Edward N. Zalta, 2019.

Broussard, Meredith. *Artificial Unintelligence: How Computers Misunderstand the World*. Mit Press. MIT Press, 2019.

Bryson, Joanna J. «Europe Is in Danger of Using the Wrong Definition of AI». *WIRED*, 2022.

Carabantes, Manuel. «Black-box artificial intelligence: an epistemological and critical analysis». *AI & SOCIETY* 35, n.º 2 (2020): 309-17. https://doi.org/10.1007/s00146-019-00888-w.

Cavero Garcés, Ignacio. «La cólera de Ludd y Swing. El luddismo industrial y agrario en el primer tercio del siglo XIX». Universidad de Zaragoza, 2020.

Comisión Europea. Libro Blanco sobre la inteligencia artificial (2020).

Consejo de Europa. Consultative Committee of Convention 108, y Council of Europe. «Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data». Strasbourg, 2017.

Cotino, Lorenzo. «Big data e inteligencia artificial . Una aproximación a su tratamiento jurídico desde los derechos fundamentales». *Dilemata* 24 (2017): 131-50.

Cowls, Josh. «'AI for Social Good': Whose Good and Who's Good? Introduction to the Special Issue on Artificial Intelligence for Social Good». *Philosophy & Technology* 34, n.º 1 (2021): 1-5. https://doi.org/10.1007/s13347-021-00466-3.

Dijck, José van. «Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology». *Surveillance and society* 12, n.º 2 (2014): 197-208. https://doi.org/https://doi.org/10.24908/ss.v12i2.4776.

Domingos, Pedro. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York, NY, USA: Basic Books, Inc., 2018.

Dreyfus, Hubert. *No Alchemy and Artificial Intelligence*. RAND, 1965.

European Group on Ethics in Science and New Technologies. «Future of Work, Future of Society», 2018. https://doi.org/doi:10.2777/029088.

Eyert, Florian, Florian Irgmaier, y Lena Ulbricht. «Extending the framework of algorithmic regulation. The Uber case». *Regulation & Governance* n/a, n.º n/a (2 de noviembre de 2020). https://doi.org/https://doi.org/10.1111/rego.12371.

Floridi, L, y F Cabitza. *L'intelligenza artificiale. L'uso delle nuove macchine*. Bompiani, 2021.

Gellert, Raphaël. «Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms?» *Regulation & Governance* n/a, n.º n/a (23 de julio de 2020). https://doi.org/https://doi.org/10.1111/rego.12349.

Green, Ben. «Data Science as Political Action: Grounding Data Science in a Politics of Justice». *Journal of Social Computing* 2, n.º 3 (2021). https://doi.org/10.23919/JSC.2021.0029.

Hagendorff, Thilo, y Katharina Wezel. «15 challenges for AI: or what AI (currently) can't do». *AI & SOCIETY* 35, n.º 2 (2020): 355-65. https://doi.org/10.1007/s00146-019-00886-y.

Hildebrandt, Mireille. «Code-driven Law: Freezing the Future and Scaling the Past». En *Is Law*

Computable?, 2020. https://doi.org/10.5040/9781509937097.ch-003.

Hoffmann, Anna Lauren. «Making Data Valuable: Political, Economic, and Conceptual Bases of Big Data». *Philosophy & Technology* 31, n.º 2 (junio de 2018): 209-12. https://doi.org/10.1007/s13347-017-0295-x.

Jaume-Palasi, Lorena, y Matthias Spielkamp. «Ethics and algorithmic processes for decision making and decision support», 2017.

Kamarinou, Dimitra, Christopher Millard, y Jatinder Singh. «Machine Learning with Personal Data». *Queen Mary School of Law Legal Studies Research Paper 247*, 2016, 1-23.

Katz, Yarden. «Manufacturing an Artificial Intelligence Revolution». *Social Science Research Network*, 2017.

Katzenbach, Christian, y Lena Ulbricht. «Algorithmic governance». *INTERNET POLICY REVIEW Journal on internet regulation* 8, n.º 4 (2019).

Korff, Douwe. «Comments on Selected Topics in the Draft EU Data Protection Regulation». *SSRN Electronic Journal*, 2012. https://doi.org/10.2139/ssrn.2150145.

Kuner, Christopher, Fred H Cate, Orla Lynskey, Christopher Millard, Nora Ni Loideain, y Dan Jerker B Svantesson. «Expanding the artificial intelligence-data protection debate». *International Data Privacy Law* 8, n.º 4 (1 de noviembre de 2018): 289-92. https://doi.org/10.1093/idpl/ipy024.

Kurzweil, Ray. *The age of inteligent machines*. Cambridge: MIT Press, 1990.

la Mata, Norberto Javier de, y Desirée Barinas Ubiñas. «La privacidad en el diseño y el diseño de la privacidad, también desde el Derecho Penal». *Eguzkilore* 28 (2014): 253-74.

Loi, Michele, y Markus Christen. «Two Concepts of Group Privacy». *Philosophy & Technology* 33, n.º 2 (mayo de 2019): 207-24. https://doi.org/10.1007/s13347-019-00351-0.

Malik, Momin M. «A Hierarchy of Limitations in Machine Learning». *CoRR* abs/2002.0 (2020).

Marcus, Gary. «Deep Learning: A Critical Appraisal». *CoRR* abs/1801.0 (2018).

Mayer-Schönberger, V, y K Cukier. *Big Data: A Revolution that We Transform How We Live, and Think*. Londres: John Murray Publishers, 2013.

McQuillan, Dan. «Algorithmic paranoia and the convivial alternative». *Big Data & Society* 3, n.º 2 (2016): 2053951716671340. https://doi.org/10.1177/2053951716671340.

———. «Data Science as Machinic Neoplatonism». *Philosophy and Technology* 31, n.º 2 (2018): 253-72. https://doi.org/10.1007/s13347-017-0273-3.

———. «The Political Affinities of AI». En *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*, editado por Andreas Sudmann, 163-74. transcript Verlag, 2020. https://doi.org/doi:10.1515/9783839447192-010.

Mejias, Ulises A., y Nick Couldry. «Datafication». *Internet Policy Review* 8, n.º 4 (2019): 1-10.

Mendez, Jesus. *Ciencia sin ficción. Cinco historias*. Penguin Random House, 2019.

Metz, Cade. «A new way for machines to see, taking shape in Toronto». *New York Times*, 2017.

Miguel, Iñigo de, y Antonio Dieguez. «¿Explicar o predecir?» *Investigación y Ciencia*, 2021.

Mitchell, Melanie. «What Does It Mean for AI to Understand?» *Quantamagazine*, 2021.

———. «Why AI is Harder than We Think». En *Proceedings of the Genetic and Evolutionary Computation Conference*, 3. GECCO '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3449639.3465421.

Müller, Vincent C. «Ethics of Artificial Intelligence and Robotics». En *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*, editado por Edward N. Zalta, 2021.

Musa Giuliano, Roberto. «Echoes of myth and magic in the language of Artificial Intelligence». *AI and Society* 35, n.º 4 (2020). https://doi.org/10.1007/s00146-020-00966-4.

Nardi, Bonnie, Bill Tomlinson, Donald J. Patterson, Jay Chen, Daniel Pargman, Barath Raghavan, y Birgit Penzenstadler. «Computing within limits». *Communications of the ACM* 61, n.º 10 (2018). https://doi.org/10.1145/3183582.

Noble, Safiya Umoja. *Algorithms of Oppression. How Search Engines Reinforce Racism*. NYU Press, 2018.

Noorman, Merel. «Computing and Moral Responsibility». En *The Stanford Encyclopedia of Philosophy (Spring 2020 Edition)*, 2020.

O'neil, Cathy. *Armas de destrucción matemática*. Editado por Traducción: Violeta Arranz de la Torre. Capitán Swing Libros, S.L., 2018.

Obermeyer, Ziad, y Ezekiel J Emanuel. «Predicting the Future — Big Data, Machine Learning, and Clinical Medicine». *New England Journal of Medicine* 375, n.º 13 (2016): 1216-19. https://doi.org/10.1056/NEJMp1606181.

Poquet Catala, Raquel. «Cuarta revolución industrial, automatización y afectación sobre la continuidad de la relación laboral». *AIS: Ars Iuris Salmanticensis* 8, n.º 1 (2020).

Prainsack, Barbara. «The political economy of digital data: introduction to the special issue». *Policy Studies* 41, n.º 5 (2020). https://doi.org/10.1080/01442872.2020.1723519.

Río Solá, María Lourdes del, José María López Santos, y Carlos Vaquero Puerta. «La

inteligencia artificial en el ámbito médico». *Revista española de investigaciones quirúrgica* 21, n.º 3 (2018): 113-16.

Rossi, Arianna, Regis Chatellier, Stefano Leucci, Rossana Ducato, y Estelle Hary. «What if data protection embraced foresight and speculative design?» En *Design Research Society (DSN)*, 1-19, 2022.

Rowson, Jonathan. «Review of "Deep Thinking"». *New in Chess*, 2017.

Santoni de Sio, Filippo, Txai Almeida, y Jeroen van den Hoven. «The future of work: freedom, justice and capital in the age of artificial intelligence». *Critical Review of International Social and Political Philosophy*, 13 de diciembre de 2021, 1-25. https://doi.org/10.1080/13698230.2021.2008204.

Sloot, Bart van der, y Sascha van Schendel. «Procedural law for the data-driven society». *Information & Communications Technology Law* 30, n.º 3 (2 de septiembre de 2021): 304-32. https://doi.org/10.1080/13600834.2021.1876331.

Turing, Alan. «Computing Machinery and Intellingence». *Mind* LIX, n.º 236 (1950): 433-60. https://doi.org/10.1093/mind/LIX.236.433.

Veale, Michael, Max Van Kleek, y Reuben Binns. «Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making». En *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018. https://doi.org/10.1145/3173574.3174014.

Veliz, Carissa. «If AI Is Predicting Your Future, Are You Still Free? Part of being human is being able to defy the odds. Algorithmic prophecies undermine that». *WIRED*, 2021.

Waldrop, M. Mitchell. «News Feature: What are the limits of deep learning?» *Proceedings of the National Academy of Sciences* 116, n.º 4 (22 de enero de 2019): 1074-77. https://doi.org/10.1073/PNAS.1821594116.

Winograd, Terry. «On some contested suppositions of generative linguistics about the scientific study of language: A response to Dresher and Hornstein's on some supposed contributions of artificial intelligence to the scientific study of language». *Cognition* 5, n.º 2 (1977): 151-79. https://doi.org/https://doi.org/10.1016/0010-0277(77)90010-5.

Yeung, Karen. «Algorithmic regulation: A critical interrogation». *Regulation and Governance* 12, n.º 4 (2018): 505-23. https://doi.org/10.1111/rego.12158.

BIBLIOGRAFÍA CAPÍTULO 1. MARCO TEÓRICO DE LA TOMA DE DECISIONES AUTOMATIZADA BASADA EN LA ELABORACIÓN DE PERFILES

Adams-Prassl, Jeremias, Reuben Binns, y Aislinn Kelly-Lyth. «Directly Discriminatory Algorithms». *The Modern Law Review* n/a, n.º n/a (1 de agosto de 2022). https://doi.org/https://doi.org/10.1111/1468-2230.12759.

Adams, Thomas K B T - Parameters. «Future Warfare and the Decline of Human Decisionmaking». *Parameters* 31, n.º 4 (21 de octubre de 2001): 57+.

Agencia Española de Protección de Datos (AEPD). «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción», 2020.

———. «La K-Anonimidad como medida de la privacidad», 2019.

Almada, Marco. «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems». En *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 2-11. ICAIL '19. New York, NY, USA: ACM, 2019. https://doi.org/10.1145/3322640.3326699.

Ananny, Mike, y Kate Crawford. «Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability». *New Media & Society* 20, n.º 3 (13 de diciembre de 2018): 973-89. https://doi.org/10.1177/1461444816676645.

Andrejevic, Mark. «Shareable and un-sharable knowledge». *Big Data & Society* 7, n.º 1 (1 de enero de 2020): 2053951720933917. https://doi.org/10.1177/2053951720933917.

Arroyo Jiménez, Luis. «Algoritmos y reglamentos». *Almacén de Derecho*, 2020.

Barocas, Solon, y Andrew D Selbst. «Big Data's Disparate Impact». *California Law Review* 104, n.º 3 (2016): 671-732.

Benson, Buster. «Cognitive bias cheat sheet. An organized list of cognitive biases because thinking is hard». *Medium*, 2016.

Bilbao Ubillos, Juan María. «Prohibición de discriminación y relaciones entre particulares». *Teoría y Realidad Constitucional*, n.º 18 (2006). https://doi.org/10.5944/trc.18.2006.6723.

Boix, Andrés. «Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones». *Revista de Derecho Público: Teoría y método* 1 (2020): 223-70. https://doi.org/https://doi.org/10.37417/RPD/vol_1_2020_33.

Bossen, Claus, Kathleen H Pine, Federico Cabitza, Gunnar Ellingsen, y Enrico Maria Piras. «Data work in healthcare: An Introduction». *Health Informatics Journal* 25, n.º 3 (12 de agosto de 2019): 465-74. https://doi.org/10.1177/1460458219864730.

Bovens, Mark. «Analysing and Assessing Accountability: A Conceptual Framework». *European Law Journal* 13, n.º 4 (1 de julio de 2007): 447-68. https://doi.org/https://doi.org/10.1111/j.1468-0386.2007.00378.x.

Brennan-Marquez, Kiel, Karen Levy, y Daniel Susser. «Strange Loops: Apparent Versus Actual Human Involvement in Automated Decision Making». *Berkeley Technology Law Journal* 34, n.º 3 (2019): 745-72.

Brkan, Maja, y Grégory Bonnet. «Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas». *European Journal of Risk Regulation* 11, n.º 1 (2020): 18-50. https://doi.org/DOI: 10.1017/err.2020.10.

Büchi, Moritz, Eduard Fosch-Villaronga, Christoph Lutz, Aurelia Tamò-Larrieux, Shruthi Velidi, y Salome Viljoen. «The chilling effects of algorithmic profiling: Mapping the issues». *Computer Law & Security Review*, 22 de noviembre de 2019, 1-15. https://doi.org/10.1016/J.CLSR.2019.105367.

Burns, Alex P. «Where is the evidence for automated triage apps?» *BMJ* 360 (2018). https://doi.org/10.1136/bmj.k885.

Burrell, Jenna. «How the machine 'thinks': Understanding opacity in machine learning algorithms». *Big Data & Society* 3, n.º 1 (2016): 1-12. https://doi.org/10.1177/2053951715622512.

Cabitza, Federico, Andrea Campagner, y Clara Balsano. «Bridging the "Last Mile" Gap between AI Implementation and Operation: "Data Awareness" That Matters». *Annals of Translational Medicine* 8, n.º 7 (abril de 2020): 501. https://doi.org/10.21037/atm.2020.03.63.

Cabitza, Federico, Andrea Campagner, y Davide Ciucci. «New Frontiers in Explainable AI: Understanding the GI to Interpret the GO BT - Machine Learning and Knowledge Extraction». En *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, editado por Andreas Holzinger, Peter Kieseberg, A Min Tjoa, y Edgar Weippl, 27-47. Cham: Springer International Publishing, 2019.

Cabitza, Federico, Andrea Campagner, Felipe Soares, Luis García de Guadiana-Romualdo, Feyissa Challa, Adela Sulejmani, Michela Seghezzi, y Anna Carobene. «The importance of being external. methodological insights for the external validation of machine learning models in medicine». *Computer Methods and Programs in Biomedicine* 208 (2021): 106288. https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106288.

Carabantes, Manuel. «Black-box artificial intelligence: an epistemological and critical

analysis». *AI & SOCIETY* 35, n.º 2 (2020): 309-17. https://doi.org/10.1007/s00146-019-00888-w.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, y Noemie Elhadad. «Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission». En *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721-30. KDD '15. New York, NY, USA: ACM, 2015. https://doi.org/10.1145/2783258.2788613.

Castelluccia, Claude, y Daniel Le Métayer. «Understanding algorithmic decision-making: Opportunities and challenges». Bruselas, 2019. https://doi.org/10.2861/536131.

Castillo Parrilla, José Antonio. «El turismo en la economía de los datos y la economía de plataformas en la UE». *Revista de Privacidad y Derecho Digital* 5, n.º 19 (2020): 115-55.

Cirillo, Davide, y Maria Jose Rementeria. «Bias and fairness in machine learning and artificial intelligence». En *Sex and Gender Bias in Technology and Artificial Intelligence*, editado por Davide Cirillo, Silvina Solarz, y Emre Guney, 3-21. Academic Press, 2022.

Cobbe, Jennifer, Michelle Seng Ah Lee, y Jatinder Singh. «Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems». En *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 598–609. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445921.

Cobbe, Jennifer, y Jatinder Singh. «Reviewable Automated Decision-Making». *Computer Law & Security Review* 39 (2020): 105475. https://doi.org/https://doi.org/10.1016/j.clsr.2020.105475.

Comisión Europea. Libro Blanco sobre la inteligencia artificial (2020).

Cranor, Lorrie Faith. «A Framework for Reasoning about the Human in the Loop». En *Proceedings of the 1st Conference on Usability, Psychology, and Security*. UPSEC'08. USA: USENIX Association, 2008.

Danks, David, y Alex John London. «Algorithmic Bias in Autonomous Systems». En *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4691–4697. IJCAI'17. AAAI Press, 2017. https://doi.org/10.24963/ijcai.2017/654.

Djeffal, Christian. «AI, Democracy and the Law». En *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*, editado por Andreas Sudmann, 255-84. transcript Verlag, 2020. https://doi.org/doi:10.1515/9783839447192-016.

Enarsson, Therese, Lena Enqvist, y Markus Naarttijärvi. «Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts». *Information & Communications Technology Law* 31, n.º 1 (2 de enero de 2022): 123-53. https://doi.org/10.1080/13600834.2021.1958860.

European Group on Ethics in Science and New Technologies. «Future of Work, Future of Society», 2018. https://doi.org/doi:10.2777/029088.

Favaretto, Maddalena, Eva De Clercq, y Bernice Simone Elger. «Big Data and discrimination: perils, promises and solutions. A systematic review». *Journal of Big Data* 6, n.º 1 (2019): 1-27. https://doi.org/10.1186/s40537-019-0177-4.

Felzmann, Heike, Eduard Fosch-Villaronga, Christoph Lutz, y Aurelia Tamò-Larrieux. «Towards Transparency by Design for Artificial Intelligence». *Science and Engineering Ethics*, 2020. https://doi.org/10.1007/s11948-020-00276-4.

Fernández López, María F. «Artículo 14 CE: Igualdad ante la Ley y prohibición de discriminación». *Diario La Ley* 9314 (2018).

Fine Licht, Karl de, y Jenny de Fine Licht. «Artificial intelligence, transparency, and public decision-making». *AI & SOCIETY*, 2020. https://doi.org/10.1007/s00146-020-00960-w.

Fischer, Joel E, Chris Greenhalgh, Wenchao Jiang, Sarvapali D Ramchurn, Feng Wu, y Tom Rodden. «In-the-loop or on-the-loop? Interactional arrangements to support team coordination with a planning agent». *Concurrency and Computation: Practice and Experience* n/a, n.º n/a (6 de marzo de 2017): e4082. https://doi.org/10.1002/cpe.4082.

Friedman, Batya, y Helen Nissenbaum. «Bias in Computer Systems». *ACM Transactions on Information Systems* 14, n.º 3 (1996): 330-47. https://doi.org/10.1145/230538.230561.

Gerards, Janneke. «The discrimination grounds of article 14 of the european convention on human rights». *Human Rights Law Review* 13, n.º 1 (2013). https://doi.org/10.1093/hrlr/ngs044.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, y Lalana Kagal. «Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning». En *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80-89, 2018.

Goodman, B, y S R Flaxman. «European Union regulations on algorithmic decision-making and a "right to explanation"». *AI Magazine* 38, n.º 3 (2017): 50-57.

Grant, Thomas D, y Damon J Wischik. «Poisonous Datasets, Poisonous Trees». En *On the path to AI: Law's prophecies and the conceptual foundations of the machine learning age*,

editado por Thomas D Grant y Damon J Wischik, 89-101. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-43582-0_8.

Green, Ben. «Data Science as Political Action: Grounding Data Science in a Politics of Justice». *Journal of Social Computing* 2, n.º 3 (2021). https://doi.org/10.23919/JSC.2021.0029.

———. «The Flaws of Policies Requiring Human Oversight of Government Algorithms». *SSRN Electronic Journal*, 2021. https://doi.org/10.2139/ssrn.3921216.

Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea (HLEG-AI). *Directrices éticas para una IA fiable*. Editado por Comisión Europea. Bruselas: Oficina de Publicaciones de la Comisión Europea, 2019. https://doi.org/doi/10.2759/14078.

Grupo de Expertos Gubernamentales sobre las tecnologías emergentes en el ámbito de los sistemas de armas autónomos letales. «Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (CCW/GGE.1/2018/3)», 2018.

Grupo de Trabajo sobre Protección de Datos del Artículo 29. «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», 2017.

Hacker, Philipp. «Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law please cite to the final version forthcoming in: Common Market Law Review». *Common Market Law Review* 55 (2018).

Hajian, Sara, y Josep Domingo-Ferrer. «Direct and Indirect Discrimination Prevention Methods». En *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, editado por Bart Custers, Toon Calders, Bart Schermer, y Tal Zarsky, 241-54. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. https://doi.org/10.1007/978-3-642-30487-3_13.

Heinrichs, Bert, y Simon B Eickhoff. «Your evidence? Machine learning algorithms for medical diagnosis and prediction». *Human Brain Mapping* 41, n.º 6 (2020): 1435-44. https://doi.org/10.1002/hbm.24886.

Hert, Paul de, y Guillermo Lazcoz. «When GDPR-principles blind each other. Accountability, not transparency, at the heart of algorithmic governance». *European Data Protection Law Review* 8, n.º 1 (2022): 1-10.

Hildebrandt, Mireille. «Primitives of Legal Protection in the Era of Data-Driven Platforms». *Georgetown Law Technology Review* 2, n.º 2 (2018): 252-73.

———. «The Issue of Bias. The Framing Powers of ML». *SSRN Electronic Journal*, 2019. https://doi.org/10.2139/ssrn.3497597.

Hoffmann, Anna Lauren. «Making Data Valuable: Political, Economic, and Conceptual Bases of Big Data». *Philosophy & Technology* 31, n.º 2 (junio de 2018): 209-12. https://doi.org/10.1007/s13347-017-0295-x.

Hohman, Fred, Minsuk Kahng, Robert Pienta, y Duen Horng Chau. «Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers». *IEEE Transactions on Visualization and Computer Graphics* 25, n.º 8 (2019). https://doi.org/10.1109/TVCG.2018.2843369.

Hond, Anne A H de, Artuur M Leeuwenberg, Lotty Hooft, Ilse M J Kant, Steven W J Nijman, Hendrikus J A van Os, Jiska J Aardoom, et al. «Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review». *npj Digital Medicine* 5, n.º 1 (2022): 2. https://doi.org/10.1038/s41746-021-00549-7.

INNOCENCE PROYECT. «A proposal for identifying and managing bias within artificial intelligence -PUBLIC COMMENT ON DRAFT NIST Special Publication 1270-», 2021.

Jacobs, Abigail Z., y Hanna M Wallach. «Measurement and fairness». En *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375-85, 2021. https://doi.org/10.1145/3442188.3445901.

Jiménez-Segovia, Reyes. «Los sistemas de armas autónomos en la Convención sobre ciertas armas convencionales: Sombras legales y eticas de una autonomía ¿bajo el control humano?» *Revista Electrónica de Estudios Internacionales*, n.º 37 (2019): 2-33. https://doi.org/10.17103/reei.37.07.

Jones, Meg Leta. «The right to a human in the loop: Political constructions of computer automation and personhood». *Social Studies of Science* 47, n.º 2 (2017): 216-39. https://doi.org/10.1177/0306312717699716.

Kahneman, Daniel, y Amos Tversky. «Subjective probability: A judgment of representativeness». *Cognitive Psychology* 3, n.º 3 (1 de julio de 1972): 430-54. https://doi.org/10.1016/0010-0285(72)90016-3.

Kiseleva, Anastasiya. «AI as a Medical Device: Is it Enough to Ensure Performance Transparency and Accountability?» *European Pharmaceutical Law Review* 4, n.º 1 (2020).

———. «Making AI's transparency transparent: notes on the EU Proposal for the AI Act». *European Law Blog*, 2021.

Kiseleva, Anastasiya, Dimitris Kotzinos, y Paul De Hert. «Transparency of AI in Healthcare as

a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations». *Frontiers in Artificial Intelligence*, 2022.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, y Sendhil Mullainathan. «Human Decisions and Machine Predictions». *The Quarterly Journal of Economics* 133, n.º 1 (2017): 237-93. https://doi.org/10.1093/qje/qjx032.

Kolkman, Daan. «The (in)credibility of algorithmic models to non-experts». *Information, Communication & Society* 25, n.º 1 (2 de enero de 2022): 93-109. https://doi.org/10.1080/1369118X.2020.1761860.

Korff, Douwe. «Comments on Selected Topics in the Draft EU Data Protection Regulation». *SSRN Electronic Journal*, 2012. https://doi.org/10.2139/ssrn.2150145.

Krupiy, Tetyana (Tanya). «A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective». *Computer Law & Security Review* 38 (2020): 105429. https://doi.org/https://doi.org/10.1016/j.clsr.2020.105429.

Laat, Paul B de. «Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?» *Philosophy & Technology* 31, n.º 4 (2018): 525-41. https://doi.org/10.1007/s13347-017-0293-z.

Lazcoz, Guillermo. «Automated decision-making under Amsterdam's District Court judgements: Drivers v. Uber and Ola». En *Time to reshape the digital society. 40th anniversary of the CRIDS*, 321-38. Larcier, 2021.

Lehr, David, y Paul Ohm. «Playing with the data: what legal scholars should learn about machine learning». *UC Davis Law Review* 51 (2017): 653-717.

Lin, Ying-Tung, Tzu-Wei Hung, y Linus Ta-Lun Huang. «Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias». *Philosophy & Technology* 34, n.º 1 (2021): 65-90. https://doi.org/10.1007/s13347-020-00406-7.

Lipton, Zachary C. «The Mythos of Model Interpretability». *Queue* 16, n.º 3 (2018): 31-57. https://doi.org/10.1145/3236386.3241340.

Mann, Monique, y Tobias Matzner. «Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination». *Big Data & Society* 6, n.º 2 (2019): 1-11. https://doi.org/10.1177/2053951719895805.

Marcus, Gary. «Deep Learning: A Critical Appraisal». *CoRR* abs/1801.0 (2018).

McQuillan, Dan. «Data Science as Machinic Neoplatonism». *Philosophy and Technology* 31, n.º 2 (2018): 253-72. https://doi.org/10.1007/s13347-017-0273-3.

Methnani, Leila, Andrea Aler Tubella, Virginia Dignum, y Andreas Theodorou. «Let Me Take Over: Variable Autonomy for Meaningful Human Control». *Frontiers in Artificial Intelligence* 4 (2021): 1-10.

Miguel Beriain, Iñigo de, Pilar Nicolás Jiménez, Maria Jose Rementeria, Davide Cirillo, Atia Cortés, Diego Saby, y Guillermo Lazcoz Moratinos. «Auditing the quality of datasets used in algorithmic decision-making systems». Bruselas, 2022. https://doi.org/10.2861/98930.

Miró Llinares, Fernando. «Inteligencia Artificial y Justicia: Más allá de los resultados lesivos causados por Robots». *Revista de Derecho Penal y Criminología* 20, n.º Jul. (2018): 87-130.

Mitchell, Melanie. «Why AI is Harder than We Think». En *Proceedings of the Genetic and Evolutionary Computation Conference*, 3. GECCO '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3449639.3465421.

Mittelstadt, Brent. «From Individual to Group Privacy in Big Data Analytics». *Philosophy & Technology* 30, n.º 4 (2017): 475-94. https://doi.org/10.1007/s13347-017-0253-7.

Mittelstadt, Brent, Chris Russell, y Sandra Wachter. «Explaining Explanations in AI». En *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279-88. FAT* '19. New York, NY, USA: ACM, 2019. https://doi.org/10.1145/3287560.3287574.

Morente Parra, Vanesa. «Big Data o el arte de analizar datos masivos. Una reflexión crítica desde los derechos fundamentales». *Derechos y libertades: Revista del Instituto Bartolomé de las Casas* 41 (2019): 225-60. https://doi.org/10.14679/1216.

Noto La Diega, Guido. «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information». *JIPITEC* 9, n.º 1 (2018).

Núñez Reiz, A., M.A. A Armengol de la Hoz, y M. Sánchez García. «Big Data Analysis y Machine Learning en medicina intensiva». *Medicina Intensiva* 43, n.º 7 (1 de octubre de 2018): 416-26. https://doi.org/10.1016/j.medin.2018.10.007.

Obregón Fernández, Aritz, y Guillermo Lazcoz Moratinos. «La supervisión humana de los sistemas de inteligencia artificial de alto riesgo. Aportaciones desde el Derecho Internacional Humanitario y el Derecho de la Unión Europea». *Revista Electrónica de Estudios Internacionales* 2021, n.º 42 (2021). https://doi.org/10.17103/reei.42.08.

Palma Ortigosa, Adrian. «Régimen jurídico de la toma de decisiones automatizadas y el uso de sistemas de inteligencia artificial en el marco del derecho a la protección de datos personales». Universidad de Valencia, 2021.

Parasuraman, Raja, y Dietrich H Manzey. «Complacency and Bias in Human Use of Automation: An Attentional Integration». *Human Factors* 52, n.º 3 (1 de junio de 2010): 381-410. https://doi.org/10.1177/0018720810376055.

Pasquale, Frank. *The Black Box Society*. *The Black Box Society*, 2015. https://doi.org/10.4159/harvard.9780674736061.

Pérez Calvo, José Luis. «Debate internacional en torno a los sistemas de armas autónomos letales. Consideraciones tecnológicas, jurídicas y éticas». *Revista general de marina* 278, n.º Abril (2020): 457-69.

Pérez Colomé, Jordi. «Por qué el despido de una investigadora negra de Google se ha convertido en un escándalo global». *El País*, 2020.

Piedmont, Ralph L. «Bias, Statistical BT - Encyclopedia of Quality of Life and Well-Being Research». editado por Alex C Michalos, 382-83. Dordrecht: Springer Netherlands, 2014. https://doi.org/10.1007/978-94-007-0753-5_2865.

Pot, Mirjam, Nathalie Kieusseyan, y Barbara Prainsack. «Not all biases are bad: equitable and inequitable biases in machine learning and radiology». *Insights into Imaging* 12, n.º 1 (2021): 13. https://doi.org/10.1186/s13244-020-00955-7.

Recht, Michael P, Marc Dewey, Keith Dreyer, Curtis Langlotz, Wiro Niessen, Barbara Prainsack, y John J Smith. «Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations». *European Radiology* 30, n.º 6 (2020): 3576-84. https://doi.org/10.1007/s00330-020-06672-5.

Rey Martínez, Fernando. «Igualdad y prohibición de discriminación: de 1978 a 2018». *Revista de Derecho Político* 100, n.º sept.-dec. (2017): 125-71.

Rodas-García, Ilene. «Alfabetizar para la democracia». *Zona Próxima* 32 (2020): 71-80.

Romeo Casabona, Carlos María. «Criminal responsability of robots and autonomous artificial intelligent systems?» *Comunicaciones en propiedad industrial y derecho de la competencia* 91, n.º Septiembre-Diciembre (2020): 167-87.

———. «El tipo del delito de acción imprudente». En *Derecho Penal. Parte General. Introducción teoría jurídica del delito*, 2ª., 133-48. Comares, 2016.

———. «Riesgo, procedimientos actuariales basados en inteligencia artificial y medidas de seguridad». *Revista Penal* 42 (2018): 165-79.

Rubio Damián, Francisco. «Automatización de la guerra: el control humano». *Ejército: de tierra español* 948 (2020): 6-11.

Sanchez Caro, Javier. «Cambio de paradigma en la relación clínico-asistencial: Aspectos

bioéticos y legales». En *E-Salud y Cambio del Modelo Sanitario*, 29-59. Fundación Merck Salud, 2020.

Schreurs, Wim, Mireille Hildebrandt, Els Kindt, y Michael Vanfleteren. «Cogitas, Ergo Sum. The Role of Data Protection Law and Non-discrimination Law in Group Profiling in the Private Sector». En *Profiling the European Citizen: Cross-disciplinary Perspectives*, editado por Mireille Hildebrandt y Serge Gutwirth, 241-64. Dordrecht: Springer, 2008.

Selbst, Andrew D., y Solon Barocas. «The Intuitive Appeal of Explainable Machines». *Fordham Law Review* 87, n.º 3 (2018): 1085-1139.

Sheridan, T B. «Human centered automation: oxymoron or common sense?» En *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, 1:823-28, 1995. https://doi.org/10.1109/ICSMC.1995.537867.

Simon, Judith, Pak-Hang Wong, y Gernot Rieder. «Algorithmic bias and the Value Sensitive Design approach». *Internet Policy Review* 9, n.º 4 (2020): 1-16. https://doi.org/10.14763/2020.4.1534.

Singh Visen, Vikram. «What is Human in the Loop Machine Learning: Why & How Used in AI?» Medium, 2020. https://medium.com/vsinghbisen/what-is-human-in-the-loop-machine-learning-why-how-used-in-ai-60c7b44eb2c0.

Sio, Filippo de, y Jeroen van den Hoven. «Meaningful Human Control over Autonomous Systems: A Philosophical Account». *Frontiers in Robotics and AI* 5 (2018): 15. https://doi.org/10.3389/frobt.2018.00015.

Skitka, Linda J, Kathleen L Mosier, y Mark Burdick. «Does automation bias decision-making?» *International Journal of Human-Computer Studies* 51, n.º 5 (1999): 991-1006. https://doi.org/https://doi.org/10.1006/ijhc.1999.0252.

Society of Automotive Engineers (SAE). «Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_201806», 2018. https://doi.org/https://doi.org/10.4271/J3016_201806.

Soriano Arnanz, Alba. «La propuesta de Reglamento de inteligencia artificial de la UE y los sistemas de alto riesgo». *Revista General de Derecho de los Sectores Regulados* 8 (2021): 1-24.

Supervisor Europeo de Protección de Datos (SEPD). «Dictamen 3/2015 -La gran oportunidad de Europa. Recomendaciones del SEPD sobre las opciones de la UE en cuanto a la reforma de la protección de datos-», 2015.

Urruela Mora, Asier. «Instrumentos de evaluación del riesgo de violencia, justicia algorítmica y

derecho penal. Perspectiva crítica». En *Libro Homenaje a Díez Ripollés*, s. f.

Valdivia, Ana, Javier Sánchez-Monedero, y Jorge Casillas. «How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness». *International Journal of Intelligent Systems* n/a, n.º n/a (1 de enero de 2021). https://doi.org/https://doi.org/10.1002/int.22354.

Veale, Michael, y Reuben Binns. «Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data». *Big Data & Society* 4, n.º 2 (2017): 1-17. https://doi.org/10.1177/2053951717743530.

Wachter, S. «Affinity profiling and discrimination by association in online behavioural advertising». *Berkeley Technology Law Journal* 35, n.º 2 (2019): 367-430.

Wachter, Sandra, Brent Daniel Mittelstadt, y Chris Russell. «Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law». *West Virginia Law Review* 123, n.º 3 (2021): 51.

Wachter, Sandra, Brent Mittelstadt, y Chris Russell. «Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI». *Computer Law and Security Review* 41 (2021). https://doi.org/10.1016/j.clsr.2021.105567.

Wagner, Ben. «Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems». *Policy & Internet* 11, n.º 1 (2019): 104-22. https://doi.org/10.1002/poi3.198.

Walmsley, Joel. «Artificial intelligence and the value of transparency». *AI & SOCIETY* 36, n.º 2 (2021): 585-95. https://doi.org/10.1007/s00146-020-01066-z.

Wellner, Galit, y Tiran Rothman. «Feminist AI: Can We Expect Our AI Systems to Become Feminist?» *Philosophy & Technology*, mayo de 2019. https://doi.org/10.1007/s13347-019-00352-z.

Xenidis, Raphaële, y Linda Senden. «EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination». En *General Principles of EU law and the EU Digital Order*, 2020.

Yeung, Karen, y Adrian Weller. «How is 'transparency' understood by legal scholars and the machine learning community?» En *BEING PROFILED*, 2019. https://doi.org/10.2307/j.ctvhrd092.9.

Zarsky, Tal. «Transparent predictions». *University of Illinois Law Review* 2013, n.º 4 (2013): 1503-70.

Zednik, Carlos. «Solving the Black Box Problem: A Normative Framework for Explainable

Artificial Intelligence». *Philosophy & Technology* 34, n.º 2 (2021): 265-88. https://doi.org/10.1007/s13347-019-00382-7.

Zerilli, John, Alistair Knott, James Maclaurin, y Colin Gavaghan. «Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?» *Philosophy & Technology* 32 (septiembre de 2019): 661–683. https://doi.org/10.1007/s13347-018-0330-6.

Zuiderveen Borgesius, Frederik J. «Strengthening legal protection against discrimination by algorithms and artificial intelligence». *The International Journal of Human Rights* 24, n.º 10 (25 de marzo de 2020): 1-22. https://doi.org/10.1080/13642987.2020.1743976.

BIBLIOGRAFÍA CAPÍTULO 2. TOMA DE DECISIONES AUTOMATIZADA EN EL RGPD: EL ARTÍCULO 22 EN LA UNIDAD DE CUIDADOS INTENSIVOS

Armada Villaverde, Maria Elena, y Ignacio Javier López Bustabad. «El reglamento general de protección de datos ante el fenómeno del "«big data"». *Revista Aranzadi de derecho y nuevas tecnologías* 51 (2019).

Bayamlıoğlu, Emre. «The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"». *Regulation & Governance*, 14 de marzo de 2021, 1-21. https://doi.org/https://doi.org/10.1111/rego.12391.

Bekkum, Marvin van, y Frederik Zuiderveen Borgesius. «Digital welfare fraud detection and the Dutch SyRI judgment». *European Journal of Social Security* 23, n.º 4 (2021). https://doi.org/10.1177/13882627211031257.

Binns, Reuben. «Human Judgment in algorithmic loops: Individual justice and automated decision-making». *Regulation & Governance*, 7 de octubre de 2020. https://doi.org/https://doi.org/10.1111/rego.12358.

Boix, Andrés. «Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones». *Revista de Derecho Público: Teoría y método* 1 (2020): 223-70. https://doi.org/https://doi.org/10.37417/RPD/vol_1_2020_33.

Brkan, Maja. «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond». *International Journal of Law and Information Technology* 27, n.º January (2019): 91-121. https://doi.org/https://doi.org/10.1093/ijlit/eay017.

Busuioc, Madalina. «Accountable Artificial Intelligence: Holding Algorithms to Account». *Public Administration Review* n/a, n.º n/a (15 de agosto de 2020). https://doi.org/https://doi.org/10.1111/puar.13293.

Bygrave, Lee A. «Minding the machine: art 15 of the EC Data Protection Directive and automated profiling». *Computer Law & Security Review* 17, n.º 1 (2001).

———. «Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making». En *Algorithmic Regulation*, editado por Karen Yeung y Martin Lodge. Oxford: Oxford University Press, 2019. https://doi.org/10.1093/oso/9780198838494.003.0011.

Demetzou, Katerina. «Data Protection Impact Assessment: A tool for accountability and the unclarified concept of 'high risk' in the General Data Protection Regulation». *Computer Law & Security Review* 35, n.º 6 (2019): 105342.

https://doi.org/https://doi.org/10.1016/j.clsr.2019.105342.

Edwards, L, y M Veale. «Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?» *IEEE Security & Privacy* 16, n.º 3 (2018): 46-54. https://doi.org/10.1109/MSP.2018.2701152.

Garriga Domínguez, Ana. «La elaboración de perfiles y su impacto en los derechos fundamentales. Una primera aproximación a su regulación en el reglamento general de protección de datos de la Unión Europea». *Derechos y libertades: Revista del Instituto Bartolomé de las Casas* 38 (2018): 107-39. https://doi.org/10.14679/1058.

Gellert, Raphaël. «Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms?» *Regulation & Governance* 16, n.º 1 (1 de enero de 2022): 156-76. https://doi.org/https://doi.org/10.1111/rego.12349.

Gil González, Elena. «Aproximación al estudio de las decisiones automatizadas en el seno del Reglamento General Europeo de Protección de Datos a la luz de las tecnologías big data y de aprendizaje computacional». *Revista española de la transparencia* 5 (2017): 165-79.

Gil González, Elena, y Paul de Hert. «Understanding the legal provisions that allow processing and profiling of personal data—an analysis of GDPR provisions and principles». *ERA Forum* 19, n.º 4 (2019): 597-621. https://doi.org/10.1007/s12027-018-0546-z.

Gillis, Talia B., y Joshua Simons. «Explanation Justification: GDPR and the Perils of Privacy». *Journal of Law and Innovation*, n.º 2 (2019): 71-99.

Grupo de Trabajo sobre Protección de Datos del Artículo 29. «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679». Bruselas, 2018.

———. «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», 2017.

Gutwirth, Serge, y Paul De Hert. «Regulating Profiling in a Democratic Constitutional State». En *Profiling the European Citizen: Cross-Disciplinary Perspectives*, editado por Mireille Hildebrandt y Serge Gutwirth, 271-302. Dordrecht: Springer Netherlands, 2008. https://doi.org/10.1007/978-1-4020-6914-7_14.

Guzman Fluja, Vicente C. «Proceso penal y justicia automatizada». *Revista General de Derecho Procesal* 53 (2021): 1-40.

Hert, Paul de, y Vagelis Papakonstantinou. «Framing Big Data in the Council of Europe and the

EU data protection law systems: Adding ‘should’ to ‘must’ via soft law to address more than only individual harms». *Computer Law & Security Review* 40 (2021): 105496. https://doi.org/https://doi.org/10.1016/j.clsr.2020.105496.

Hildebrandt, Mireille. «Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning». *Theoretical Inquiries in Law* 20, n.º 1 (2019): 83. https://doi.org/10.1515/til-2019-0004.

———. «Technology and the End of Law». En *Facing the limits of the law*, editado por Erik Claes, Wouter Devroe, y Bert Keirsbilck, 443-64. Springer, 2009. https://doi.org/10.1007/978-3-540-79856-9.

Hilden, Jockum. «The Politics of Datafication: The influence of lobbyists on the EU’s data protection reform and its consequences for the legitimacy of the General Data Protection Regulation». University of Helsinki, 2019.

Huq, Aziz Z. «A Right to a Human Decision». *Virginia Law Review* 106, n.º 3 (2020): 611-88.

Janssen, Heleen L. «An approach for a fundamental rights impact assessment to automated decision-making». *International Data Privacy Law* 10, n.º 1 (1 de febrero de 2020): 76-106. https://doi.org/10.1093/idpl/ipz028.

Jones, Meg Leta. «The right to a human in the loop: Political constructions of computer automation and personhood». *Social Studies of Science* 47, n.º 2 (2017): 216-39. https://doi.org/10.1177/0306312717699716.

Jones, Meg Leta, y Elizabeth Edenberg. «Troubleshooting AI and Consent». En *The Oxford Handbook of Ethics of AI*, editado por Markus D. Dubber, Frank Pasquale, y Das Sunit, 359-74. Oxford University Press, 2020. https://doi.org/10.1093/oxfordhb/9780190067397.013.23.

Jove, Daniel. «Quo vadis, intimidad?» En *Setenta años de Constitución Italiana y cuarenta años de Constitución Española*, 151-66. CEPC, 2020.

Kamarinou, Dimitra, Christopher Millard, y Jatinder Singh. «Machine Learning with Personal Data». *Queen Mary School of Law Legal Studies Research Paper 247*, 2016, 1-23.

Korff, Douwe. «Comments on Selected Topics in the Draft EU Data Protection Regulation». *SSRN Electronic Journal*, 2012. https://doi.org/10.2139/ssrn.2150145.

la Mata, Norberto Javier de, y Desirée Barinas Ubiñas. «La privacidad en el diseño y el diseño de la privacidad, también desde el Derecho Penal». *Eguzkilore* 28 (2014): 253-74.

Lazcoz, Guillermo. «Automated decision-making under Amsterdam’s District Court judgements: Drivers v. Uber and Ola». En *Time to reshape the digital society. 40th*

anniversary of the CRIDS, 321-38. Larcier, 2021.

Lynskey, Orla. «Deconstructing data protection: The "added-value" of a right to data protection in the eu legal order». *International and Comparative Law Quarterly* 63, n.º 3 (2014). https://doi.org/10.1017/S0020589314000244.

Malgieri, Gianclaudio. «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations». *Computer Law & Security Review*, 2019. https://doi.org/https://doi.org/10.1016/j.clsr.2019.05.002.

Malgieri, Gianclaudio, y Giovanni Comandé. «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation». *International Data Privacy Law* 7, n.º 4 (1 de noviembre de 2017): 243-65. https://doi.org/10.1093/idpl/ipx019.

Malgieri, Gianclaudio, y Jędrzej Niklas. «Vulnerable data subjects». *Computer Law & Security Review* 37 (2020): 105415. https://doi.org/https://doi.org/10.1016/j.clsr.2020.105415.

Mendoza, Isak, y Lee A Bygrave. «The Right Not to be Subject to Automated Decisions Based on Profiling BT -». En *EU Internet Law: Regulation and Enforcement*, editado por Tatiana-Eleni Synodinou, Philippe Jougleux, Christiana Markou, y Thalia Prastitou, 77-98. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-64955-9_4.

Noto La Diega, Guido. «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information». *JIPITEC* 9, n.º 1 (2018).

Palma Ortigosa, Adrian. «Automated Decision-Making in the GDPR. Algorithms in the Scope of the Data Protection». *Revista General De Derecho Administrativo*, n.º 50 (2019).

Rallo Lombarte, Artemi. «El nuevo derecho de protección de datos». *Revista Española de Derecho Constitucional*, n.º 116 (2019). https://doi.org/10.18042/cepc/redc.116.02.

Roig, Antoni. *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*. Barcelona: Bosch Editor, 2020.

Sancho Lopez, Marina. «Legal Strategies To Ensure Fundamental Rights in Front of the Chanllenges of Big Data». *Revista General De Derecho Administrativo*, n.º 50 (2019): 1-28.

Sartor, Giovanni, y Francesca Lagioia. «The impact of the General Data Protection Regulation (GDPR) on artificial intelligence (PE 641.530)», 2020.

Schermer, Bart W. «The limits of privacy in automated profiling and data mining». *Computer*

Law & Security Review 27, n.º 1 (1 de febrero de 2011): 45-52. https://doi.org/10.1016/J.CLSR.2010.11.009.

Selbst, Andrew D, y Julia Powles. «Meaningful information and the right to explanation». *International Data Privacy Law* 7, n.º 4 (1 de noviembre de 2017): 233-42. https://doi.org/10.1093/idpl/ipx022.

Sloot, Bart van der. «The Quality of Law: How the European Court of Human Rights gradually became a European Constitutional Court for privacy cases». *JIPITEC* 11, n.º 2 (2020): 160-85.

Solove, Daniel J. «Privacy and Power: Computer Databases and Metaphors for Information Privacy». *Stanford Law Review* 53 (2001): 1393-1462.

———. *The Digital Person: Technology and Privacy in the Information Age*. New York University Press, 2004.

Supervisor Europeo de Protección de Datos (SEPD). «Dictamen 3/2015 -La gran oportunidad de Europa. Recomendaciones del SEPD sobre las opciones de la UE en cuanto a la reforma de la protección de datos-», 2015.

Tosoni, Luca. «The right to object to automated individual decisions: resolving the ambiguity of Article 22(1) of the General Data Protection Regulation». *International Data Privacy Law*, 14 de enero de 2021. https://doi.org/10.1093/idpl/ipaa024.

Troncoso Reigada (Dir.), Antonio. *Comentario al Reglamento General de Protección de Datos y a la Ley Orgánica de Protección de Datos personales y Garantía de los Derechos Digitales*. Editado por Antonio Troncoso Reigada. Thomson Reuters-Aranzadi, 2021.

Turégano Mansilla, Isabel. «Los valores detrás de la privacidad». *Doxa. Cuadernos de Filosofía del Derecho* 43 (2020): 255-83. https://doi.org/10.14198/DOXA2020.43.10.

Veale, Michael, y Lilian Edwards. «Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling». *Computer Law & Security Review* 34, n.º 2 (1 de abril de 2018): 398-404. https://doi.org/10.1016/J.CLSR.2017.12.002.

Wachter, S. «Affinity profiling and discrimination by association in online behavioural advertising». *Berkeley Technology Law Journal* 35, n.º 2 (2019): 367-430.

Wachter, Sandra, y Brent Mittelstadt. «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI». *Colum. Bus. L. Rev.*, n.º 1 (2019): 1-130.

Wachter, Sandra, Brent Mittelstadt, y Luciano Floridi. «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation».

International Data Privacy Law 7, n.º 2 (1 de mayo de 2017): 76-99. https://doi.org/10.1093/idpl/ipx005.

Warren, Samuel D., y Louis D. Brandeis. «The right to privacy». *Harvard Law Review* IV, n.º 5 (1890): 194-220.

Yeung, Karen, y Lee A. Bygrave. «Demystifying the modernized European data protection regime: Cross-disciplinary insights from legal and regulatory governance scholarship». *Regulation and Governance* 16, n.º 1 (2022). https://doi.org/10.1111/rego.12401.

Zarsky, Tal. «Incompatible: The GDPR in the Age of Big Data». *Seton Hall Law Review* 47, n.º 4 (2017).

Zuiderveen Borgesius, Frederik J. «Strengthening legal protection against discrimination by algorithms and artificial intelligence». *The International Journal of Human Rights* 24, n.º 10 (25 de marzo de 2020): 1-22. https://doi.org/10.1080/13642987.2020.1743976.

BIBLIOGRAFÍA CAPÍTULO 3. TRES PILARES SOBRE LOS QUE INTERPRETAR Y HACER EFECTIVA LA REGULACIÓN DE LA TOMA DE DECISIONES EN EL RGPD: DERECHO A LA INTERVENCIÓN HUMANA, DERECHO A LA INFORMACIÓN Y DERECHO A IMPUGNAR LA DECISIÓN. UNA PROPUESTA TERAPÉUTICA

Agencia Española de Protección de Datos (AEPD). «Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción», 2020.

Almada, Marco. «Automated Decision-Making as a Data Protection Issue», 2021.

———. «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems». En *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 2-11. ICAIL '19. New York, NY, USA: ACM, 2019. https://doi.org/10.1145/3322640.3326699.

Armada Villaverde, Maria Elena, y Ignacio Javier López Bustabad. «El reglamento general de protección de datos ante el fenómeno del "«big data"». *Revista Aranzadi de derecho y nuevas tecnologías* 51 (2019).

Barocas, Solon, Andrew D Selbst, y Manish Raghavan. «The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons». En *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 80–89. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3351095.3372830.

Bayamlıoğlu, Emre. «The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation"». *Regulation & Governance*, 14 de marzo de 2021, 1-21. https://doi.org/https://doi.org/10.1111/rego.12391.

Berscheid, Janelle, y Francois Roewer-Despres. «Beyond Transparency: A Proposed Framework for Accountability in Decision-Making AI Systems». *AI Matters* 5, n.º 2 (2019): 13-22.

Binns, Reuben. «Human Judgment in algorithmic loops: Individual justice and automated decision-making». *Regulation & Governance*, 7 de octubre de 2020. https://doi.org/https://doi.org/10.1111/rego.12358.

Brennan-Marquez, Kiel, y Stephen Henderson. «Artificial Intelligence and Role-Reversible Judgment». *Journal of Criminal Law and Criminology* 109, n.º 2 (1 de enero de 2019).

Brkan, Maja. «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond». *International Journal of Law and Information Technology* 27,

n.º January (2019): 91-121. https://doi.org/https://doi.org/10.1093/ijlit/eay017.

Brkan, Maja, y Grégory Bonnet. «Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas». *European Journal of Risk Regulation* 11, n.º 1 (2020): 18-50. https://doi.org/DOI: 10.1017/err.2020.10.

Bygrave, Lee A. «Minding the machine: art 15 of the EC Data Protection Directive and automated profiling». *Computer Law & Security Review* 17, n.º 1 (2001).

Cobbe, Jennifer, y Jatinder Singh. «Reviewable Automated Decision-Making». *Computer Law & Security Review* 39 (2020): 105475. https://doi.org/https://doi.org/10.1016/j.clsr.2020.105475.

Comisión Europea. Libro Blanco sobre la inteligencia artificial (2020).

Consejo de Europa. Consultative Committee of Convention 108, y Council of Europe. «Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data». Strasbourg, 2017.

Cotino, Lorenzo. «Ética en el diseño y confiable para el desarrollo de la robótica, inteligencia artificial y el big data y su utilidad desde el derecho». *Revista Catalana de Dret Públic* 58 (2019): 29-48.

Crabtree, Andy, Lachlan Urquhart, y Jiahong Chen. «Right to an Explanation Considered Harmful». *SSRN Electronic Journal*, 2019. https://doi.org/10.2139/ssrn.3384790.

Edwards, L, y M Veale. «Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?» *IEEE Security & Privacy* 16, n.º 3 (2018): 46-54. https://doi.org/10.1109/MSP.2018.2701152.

Edwards, Lilian, y Michael Veale. *Slave to the Algorithm? Why a Right to Explanationn is Probably Not the Remedy You are Looking for*. *Duke L. & Tech. Rev.* Vol. 18, 2017. https://doi.org/10.2139/ssrn.2972855.

Favaretto, Maddalena, Eva De Clercq, y Bernice Simone Elger. «Big Data and discrimination: perils, promises and solutions. A systematic review». *Journal of Big Data* 6, n.º 1 (2019): 1-27. https://doi.org/10.1186/s40537-019-0177-4.

Gacutan, Joshua, y Niloufer Selvadurai. «A statutory right to explanation for decisions generated using artificial intelligence». *International Journal of Law and Information Technology* 28, n.º 3 (1 de septiembre de 2020): 193-216. https://doi.org/10.1093/ijlit/eaaa016.

Garriga Domínguez, Ana. «La elaboración de perfiles y su impacto en los derechos

fundamentales. Una primera aproximación a su regulación en el reglamento general de protección de datos de la Unión Europea». *Derechos y libertades: Revista del Instituto Bartolomé de las Casas* 38 (2018): 107-39. https://doi.org/10.14679/1058.

Geburczyk, Filip. «Automated administrative decision-making under the influence of the GDPR – Early reflections and upcoming challenges». *Computer Law & Security Review* 41 (2021): 105538. https://doi.org/https://doi.org/10.1016/j.clsr.2021.105538.

Gil González, Elena, y Paul de Hert. «Understanding the legal provisions that allow processing and profiling of personal data—an analysis of GDPR provisions and principles». *ERA Forum* 19, n.º 4 (2019): 597-621. https://doi.org/10.1007/s12027-018-0546-z.

Grupo de Trabajo sobre Protección de Datos del Artículo 29. «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679». Bruselas, 2018.

Guzman Fluja, Vicente C. «Proceso penal y justicia automatizada». *Revista General de Derecho Procesal* 53 (2021): 1-40.

Hallinan, Dara, y Frederik Zuiderveen Borgesius. «Opinions can be incorrect (in our opinion)! On data protection law's accuracy principle». *International Data Privacy Law* 10, n.º 1 (1 de febrero de 2020): 1-10. https://doi.org/10.1093/idpl/ipz025.

Hamon, Ronan, Henrik Junklewitz, Gianclaudio Malgieri, Paul De Hert, Laurent Beslay, y Ignacio Sanchez. «Impossible Explanations? Beyond Explainable AI in the GDPR from a COVID-19 Use Case Scenario». En *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 549–559. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445917.

Hildebrandt, Mireille. «Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning». *Theoretical Inquiries in Law* 20, n.º 1 (2019): 83. https://doi.org/10.1515/til-2019-0004.

Hoepman, Jaap-Henk. «Transparency is the perfect cover-up (if the sun does not shine)». En *BEING PROFILED*, editado por EMRE BAYAMLIOĞLU, IRINA BARALIUC, LIISA JANSSENS, y MIREILLE HILDEBRANDT, 46-51. COGITAS ERGO SUM: 10 Years of Profiling the European Citizen. Amsterdam University Press, 2018. https://doi.org/10.2307/j.ctvhrd092.11.

Hoofnagle, Chris Jay, Bart van der Sloot, y Frederik Zuiderveen Borgesius. «The European Union general data protection regulation: what it is and what it means». *Information & Communications Technology Law* 28, n.º 1 (2 de enero de 2019): 65-98. https://doi.org/10.1080/13600834.2019.1573501.

Huq, Aziz Z. «A Right to a Human Decision». *Virginia Law Review* 106, n.º 3 (2020): 611-88.

ICO. «Guide to the UK General Data Protection Regulation (UK GDPR)», 2020.

Jones, Meg Leta. «The right to a human in the loop: Political constructions of computer automation and personhood». *Social Studies of Science* 47, n.º 2 (2017): 216-39. https://doi.org/10.1177/0306312717699716.

Jove, Daniel. «Peter Nowak v Data Protection Commissioner». *European Data Protection Law Review* 5, n.º 2 (2019): 175-83.

Kluttz, Daniel N, Nitin Kohli, y Deirdre K Mulligan. «Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions». En *After the Digital Tornado: Networks, Algorithms, Humanity*, editado por Kevin Werbach, 137-52. Cambridge: Cambridge University Press, 2020. https://doi.org/DOI: undefined.

Koivisto, Ida. «Thinking Inside the Box: The Promise and Boundaries of Transparency in Automated Decision-Making». *Academy of European Law working papers* 2020/01 (2020): 1-22.

Korff, Douwe. «Comments on Selected Topics in the Draft EU Data Protection Regulation». *SSRN Electronic Journal*, 2012. https://doi.org/10.2139/ssrn.2150145.

———. «New Challenges to Data Protection Study - Comparative Chart: Divergencies between Data Protection Laws in the EU». *SSRN Electronic Journal*, 2010. https://doi.org/10.2139/ssrn.1638951.

Kroll, Joshua A. «The fallacy of inscrutability». *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, n.º 2133 (28 de noviembre de 2018): 20180084. https://doi.org/10.1098/rsta.2018.0084.

Lazcoz, Guillermo. «Automated decision-making under Amsterdam's District Court judgements: Drivers v. Uber and Ola». En *Time to reshape the digital society. 40th anniversary of the CRIDS*, 321-38. Larcier, 2021.

Malgieri, Gianclaudio. «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations». *Computer Law & Security Review*, 2019. https://doi.org/https://doi.org/10.1016/j.clsr.2019.05.002.

Malgieri, Gianclaudio, y Giovanni Comandé. «Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation». *International Data Privacy Law* 7, n.º 4 (1 de noviembre de 2017): 243-65. https://doi.org/10.1093/idpl/ipx019.

Mendoza, Isak, y Lee A Bygrave. «The Right Not to be Subject to Automated Decisions Based

on Profiling BT -». En *EU Internet Law: Regulation and Enforcement*, editado por Tatiana-Eleni Synodinou, Philippe Jougleux, Christiana Markou, y Thalia Prastitou, 77-98. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-64955-9_4.

Morente Parra, Vanesa. «Big Data o el arte de analizar datos masivos. Una reflexión crítica desde los derechos fundamentales». *Derechos y libertades: Revista del Instituto Bartolomé de las Casas* 41 (2019): 225-60. https://doi.org/10.14679/1216.

Noto La Diega, Guido. «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information». *JIPITEC* 9, n.º 1 (2018).

Palma Ortigosa, Adrian. «Automated Decision-Making in the GDPR. Algorithms in the Scope of the Data Protection». *Revista General De Derecho Administrativo*, n.º 50 (2019).

Ponce Solé, Juli. «Inteligencia artificial, Derecho administrativo y reserva de humanidad: algoritmos y procedimiento administrativo debido tecnológico». *Revista General de Derecho Administrativo* 50 (2019): 1-52.

Roig, Antoni. *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*. Barcelona: Bosch Editor, 2020.

———. «Safeguards for the right not to be subject to a decision based solely on automated processing (Article 22 GDPR)». *European Journal of Law and Technology* 8, n.º 3 (2017).

Romeo Casabona, Carlos María. «Criminal responsability of robots and autonomous artificial intelligent systems?» *Comunicaciones en propiedad industrial y derecho de la competencia* 91, n.º Septiembre-Diciembre (2020): 167-87.

———. «Datos personales (Comentario al artículo 4. 1 RGPD)». En *Comentario al Reglamento General de Protección de Datos y a la Ley Orgánica de Protección de Datos personales y Garantía de los Derechos Digitales*, editado por Antonio Troncoso Reigada, 573-89, 2021.

Selbst, Andrew D., y Solon Barocas. «The Intuitive Appeal of Explainable Machines». *Fordham Law Review* 87, n.º 3 (2018): 1085-1139.

Selbst, Andrew D, y Julia Powles. «Meaningful information and the right to explanation». *International Data Privacy Law* 7, n.º 4 (1 de noviembre de 2017): 233-42. https://doi.org/10.1093/idpl/ipx022.

Supervisor Europeo de Protección de Datos (SEPD). «Dictamen 3/2015 -La gran oportunidad de Europa. Recomendaciones del SEPD sobre las opciones de la UE en cuanto a la reforma de la protección de datos-», 2015.

Turégano Mansilla, Isabel. «Los valores detrás de la privacidad». *Doxa. Cuadernos de Filosofía del Derecho* 43 (2020): 255-83. https://doi.org/10.14198/DOXA2020.43.10.

Veale, Michael, y Lilian Edwards. «Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling». *Computer Law & Security Review* 34, n.º 2 (1 de abril de 2018): 398-404. https://doi.org/10.1016/J.CLSR.2017.12.002.

Vries, Ekaterina de. «Machine learning/Informational fundamental rights: Makings of sameness and difference». VUB, 2016.

Wachter, Sandra, y Brent Mittelstadt. «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI». *Colum. Bus. L. Rev.*, n.º 1 (2019): 1-130.

Wachter, Sandra, Brent Mittelstadt, y Luciano Floridi. «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation». *International Data Privacy Law* 7, n.º 2 (1 de mayo de 2017): 76-99. https://doi.org/10.1093/idpl/ipx005.

Wachter, Sandra, Brent Mittelstadt, y Chris Russell. «Counterfactual explanations without opening the black box: automated decisions and the GDPR VO - 31 RT - Journal Article». *Harvard Journal of Law and Technology* 31, n.º 2 (2018): 841-87.

Wagner, Ben. «Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems». *Policy & Internet* 11, n.º 1 (2019): 104-22. https://doi.org/10.1002/poi3.198.

Zuiderveen Borgesius, Frederik J. «Strengthening legal protection against discrimination by algorithms and artificial intelligence». *The International Journal of Human Rights* 24, n.º 10 (25 de marzo de 2020): 1-22. https://doi.org/10.1080/13642987.2020.1743976.

BIBLIOGRAFÍA CAPÍTULO 4. LA INTERVENCIÓN HUMANA Y EL PRINCIPIO DE RESPONSABILIDAD EN EL TRATAMIENTO DE DATOS PERSONALES: UN ENFOQUE BASADO EN LA EVIDENCIA A TRAVÉS DE LA EVALUACIÓN DE IMPACTO. UNA PROPUESTA DESDE LA MEDICINA PREVENTIVA

Agencia Española de Protección de Datos (AEPD). «Gestión del riesgo y evaluación de impacto en tratamientos de datos personales», 2021.

Almada, Marco. «Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems». En *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 2-11. ICAIL '19. New York, NY, USA: ACM, 2019. https://doi.org/10.1145/3322640.3326699.

Berendt, Bettina, y Sören Preibusch. «Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop-and Under the Looking Glass». *Big Data* 5, n.º 2 (2017): 135-52. https://doi.org/doi:10.1089/big.2016.0055.

Binns, Reuben. «Human Judgment in algorithmic loops: Individual justice and automated decision-making». *Regulation & Governance*, 7 de octubre de 2020. https://doi.org/https://doi.org/10.1111/rego.12358.

Bovens, Mark. «Analysing and Assessing Accountability: A Conceptual Framework». *European Law Journal* 13, n.º 4 (1 de julio de 2007): 447-68. https://doi.org/https://doi.org/10.1111/j.1468-0386.2007.00378.x.

Brennan-Marquez, Kiel, y Stephen Henderson. «Artificial Intelligence and Role-Reversible Judgment». *Journal of Criminal Law and Criminology* 109, n.º 2 (1 de enero de 2019).

Brkan, Maja. «Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond». *International Journal of Law and Information Technology* 27, n.º January (2019): 91-121. https://doi.org/https://doi.org/10.1093/ijlit/eay017.

Cabitza, Federico. «Many say that AI can outperform human doctors. Is it true?» LinkedIn, 2018. https://www.linkedin.com/pulse/many-say-ai-can-outperform-human-doctors-true-federico-cabitza/.

Cabitza, Federico, Andrea Campagner, y Edoardo Datteri. «To Err is (only) Human. Reflections on How to Move from Accuracy to Trust for Medical AI». En *ITAIS 2020, the XVII Conference of the Italian Chapter of AIS Organizing in a digitized world: Diversity, Equality and Inclusion*, editado por Federica Ceci, Andrea Prencipe, y Paolo Spagnoletti, 36-49. Cham: Springer International Publishing, 2021.

Cabitza, Federico, Andrea Campagner, y Luca Maria Sconfienza. «Studying human-AI collaboration protocols: the case of the Kasparov's law in radiological double reading».

Health Information Science and Systems 9, n.º 1 (2021): 8. https://doi.org/10.1007/s13755-021-00138-8.

Casanova Asencio, Andrea Salud. «Mecanismos de prevención del acceso indebido a la historia clínica por parte del personal sanitario y nueva legislación de protección de datos». *Bioderecho.es*, n.º 7 (31 de marzo de 2018). https://doi.org/10.6018/bioderecho.360771.

Citron, Danielle Keats. «Technological Due Process». *Washington Law Review* 85, n.º 6 (2008): 1249-1313.

Cobbe, Jennifer, Michelle Seng Ah Lee, y Jatinder Singh. «Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems». En *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 598–609. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445921.

Cobbe, Jennifer, y Jatinder Singh. «Reviewable Automated Decision-Making». *Computer Law & Security Review* 39 (2020): 105475. https://doi.org/https://doi.org/10.1016/j.clsr.2020.105475.

Comité Europeo de Protección de Datos (CEPD). «Opinion 12/2019 on the draft list of the com supervisory authority of Spain petent regarding the processing operations protection exempt from the requirement of a data impact assessment (Article 35 (5) GDPR)», 2019.

Comité Europeo de Protección de Datos (CEPD), y Supervisor Europeo de Protección de Datos (SEPD). «Dictamen conjunto 5/2021 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)», 2021.

Cotino, Lorenzo. «Ética en el diseño y confiable para el desarrollo de la robótica, inteligencia artificial y el big data y su utilidad desde el derecho». *Revista Catalana de Dret Públic* 58 (2019): 29-48.

———. «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* 50 (2019): 1-37.

Cotino, Lorenzo, José Antonio Castillo Parrilla, Idoia Salazar, Richard Benjamins, María Cumbreras, y Adaya María Esteban. «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*, 2021, 1-15.

Cserne, Peter, Rossana Ducato, y Patricia Živković. «Commentary to the Commission's proposal for the "AI Act" – Response to selected issues», 2021.

EDPS. «Opinion 4/2020 on the European Commission's White Paper on Artificial Intelligence – A European approach to excellence and trust», 2020.

Edwards, L, y M Veale. «Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?» *IEEE Security & Privacy* 16, n.º 3 (2018): 46-54. https://doi.org/10.1109/MSP.2018.2701152.

Gandy, Oscar H. «Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems». *Ethics and Information Technology* 12, n.º 1 (marzo de 2010): 29-42. https://doi.org/10.1007/s10676-009-9198-6.

Gillis, Talia B., y Joshua Simons. «Explanation Justification: GDPR and the Perils of Privacy». *Journal of Law and Innovation*, n.º 2 (2019): 71-99.

Grupo de Trabajo sobre Protección de Datos del Artículo 29. «Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679». Bruselas, 2018.

———. «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», 2017.

Gstrein, Oskar Josef. «European AI Regulation: Brussels Effect versus Human Dignity?» *Forthcoming, Zeitschrift für Europarechtliche Studien (ZEuS)*, n.º 4 (2022): 1-24.

Hawath, Mariam. «Regulating Automated Decision-Making: An Analysis of Control over Processing and Additional Safeguards in Article 22 of the GDPR.» *European Data Protection Law Review* 7, n.º 2 (2021): 161-73.

Henin, Clément, y Daniel Le Métayer. «A framework to contest and justify algorithmic decisions». *AI and Ethics* 1 (2021): 463-76. https://doi.org/10.1007/s43681-021-00054-3.

Hert, Paul de, y Guillermo Lazcoz. «When GDPR-principles blind each other. Accountability, not transparency, at the heart of algorithmic governance». *European Data Protection Law Review* 8, n.º 1 (2022): 1-10.

Hildebrandt, Mireille. «Comments on White Paper on AI (EC)», 2020.

Huq, Aziz Z. «A Right to a Human Decision». *Virginia Law Review* 106, n.º 3 (2020): 611-88.

Jamieson, Trevor, y Avi Goldfarb. «Clinical considerations when applying machine learning to decision-support tasks versus automation». *BMJ Quality & Safety* 28, n.º 10 (1 de octubre de 2019): 778 LP - 781. https://doi.org/10.1136/bmjqs-2019-009514.

Janssen, Heleen L. «An approach for a fundamental rights impact assessment to automated decision-making». *International Data Privacy Law* 10, n.º 1 (1 de febrero de 2020): 76-

106. https://doi.org/10.1093/idpl/ipz028.

Kahneman, Daniel, Andrew M. Rosenfield, Linnea Gandhi, y Tom Blaser. «Noise: How to overcome the high, hidden cost of inconsistent decision making». *Harvard Business Review*, n.º October (2016).

Kahneman, Daniel, Olivier Sibony, y Carr Sunstein. *Noise: A Flaw in Human Judgment*. Little Brown Spark, 2021.

Kaminski, Margot E, y Gianclaudio Malgieri. «Algorithmic impact assessments under the GDPR: producing multi-layered explanations». *International Data Privacy Law* 11, n.º 2 (1 de abril de 2021): 125-44. https://doi.org/10.1093/idpl/ipaa020.

———. «Multi-Layered Explanations from Algorithmic Impact Assessments in the GDPR». En *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 68–79. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3351095.3372875.

Korff, Douwe. «Comments on Selected Topics in the Draft EU Data Protection Regulation». *SSRN Electronic Journal*, 2012. https://doi.org/10.2139/ssrn.2150145.

Kuner, Christopher, Dan Jerker B Svantesson, Fred H Cate, Orla Lynskey, y Christopher Millard. «Machine learning with personal data: is data protection law smart enough to meet the challenge?» *International Data Privacy Law* 7, n.º 1 (1 de febrero de 2017): 1-2. https://doi.org/10.1093/idpl/ipx003.

Malgieri, Gianclaudio. «Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations». *Computer Law & Security Review*, 2019. https://doi.org/https://doi.org/10.1016/j.clsr.2019.05.002.

———. «"Just" Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation». *Law and Business* 1, n.º 1 (2021): 16-28. https://doi.org/doi:10.2478/law-2021-0003.

Mantelero, Alessandro. «AI and Big Data: A blueprint for a human rights, social and ethical impact assessment». *Computer Law & Security Review* 34, n.º 4 (2018): 754-72. https://doi.org/https://doi.org/10.1016/j.clsr.2018.05.017.

Mantelero, Alessandro, y Maria Samantha Esposito. «An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems». *Computer Law & Security Review* 41 (2021): 105561. https://doi.org/https://doi.org/10.1016/j.clsr.2021.105561.

Martínez, Ricard. «Cuestiones de ética jurídica al abordar proyectos de Big Data. El contexto

del Reglamento general de protección de datos». *Dilemata* 24 (2017): 151-64.

Mayer-Schönberger, Viktor. «Paradigm shift». *Computer Law and Security Review* 40 (2021). https://doi.org/10.1016/j.clsr.2020.105515.

McGuire, M R. «The laughing policebot: automation and the end of policing». *Policing and Society* 31, n.º 1 (2 de enero de 2021): 20-36. https://doi.org/10.1080/10439463.2020.1810249.

Montalvo, Federico de. «¿Puede la máquina sustituir al hombre? Una reflexión jurídica sobre el ojo clínico y la responsabilidad en tiempos de Big Data». *Fronteras CTR, Revista de Ciencia Tecnología y Religión*, 2018.

Noto La Diega, Guido. «Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information». *JIPITEC* 9, n.º 1 (2018).

Palma Ortigosa, Adrian. «Automated Decision-Making in the GDPR. Algorithms in the Scope of the Data Protection». *Revista General De Derecho Administrativo*, n.º 50 (2019).

———. «Régimen jurídico de la toma de decisiones automatizadas y el uso de sistemas de inteligencia artificial en el marco del derecho a la protección de datos personales». Universidad de Valencia, 2021.

Parasuraman, Raja, y Dietrich H Manzey. «Complacency and Bias in Human Use of Automation: An Attentional Integration». *Human Factors* 52, n.º 3 (1 de junio de 2010): 381-410. https://doi.org/10.1177/0018720810376055.

Roig, Antoni. *Las garantías frente a las decisiones automatizadas. Del Reglamento General de Protección de Datos a la gobernanza algorítmica*. Barcelona: Bosch Editor, 2020.

Sandhu, Ajay, y Peter Fussey. «The 'uberization of policing'? How police negotiate and operationalise predictive policing technology». *Policing and Society* 31, n.º 1 (2 de enero de 2021): 66-81. https://doi.org/10.1080/10439463.2020.1803315.

Schwemer, Sebastian Felix, Letizia Tomada, y Tommaso Pasini. «Legal AI Systems in the EU's proposed Artificial Intelligence Act». En *Proceedings of the Second International Workshop on AI and Intelligent Assistance for Legal Professionals in the Digital Workplace*, 1-8, 2021.

Selbst, Andrew D. «Disparate Impact in Big Data Policing». *Georgia Law Review* 52 (2017): 109-95.

Soriano Arnanz, Alba. «La propuesta de Reglamento de inteligencia artificial de la UE y los sistemas de alto riesgo». *Revista General de Derecho de los Sectores Regulados* 8 (2021):

1-24.

Tamò-Larrieux, Aurelia. «Decision-making by machines: Is the ‘Law of Everything’ enough?» *Computer Law & Security Review* 41 (2021): 105541. https://doi.org/https://doi.org/10.1016/j.clsr.2021.105541.

Turégano Mansilla, Isabel. «Los valores detrás de la privacidad». *Doxa. Cuadernos de Filosofía del Derecho* 43 (2020): 255-83. https://doi.org/10.14198/DOXA2020.43.10.

Veale, Michael, y Lilian Edwards. «Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling». *Computer Law & Security Review* 34, n.º 2 (1 de abril de 2018): 398-404. https://doi.org/10.1016/J.CLSR.2017.12.002.

Veale, Michael, y Frederik Zuiderveen Borgesius. «Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach». *Computer Law Review International* 22, n.º 4 (2021): 97-112. https://doi.org/doi:10.9785/cri-2021-220402.

Wachter, Sandra, y Brent Mittelstadt. «A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI». *Colum. Bus. L. Rev.*, n.º 1 (2019): 1-130.

Wagner, Ben. «Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems». *Policy & Internet* 11, n.º 1 (2019): 104-22. https://doi.org/10.1002/poi3.198.

Yeung, Karen. «Algorithmic regulation: A critical interrogation». *Regulation and Governance* 12, n.º 4 (2018): 505-23. https://doi.org/10.1111/rego.12158.

En Madrid, a 20 de diciembre de 2022