DESIGUALDADES ALGORÍTMICAS: CONDUCTAS DE ALTO RIESGO PARA LOS DERECHOS HUMANOS*

ALGORITHMIC INEQUALITIES: HIGH-RISK PRACTICES TO HUMAN RIGHTS

María José Añón Roig
Universitat de València
https://orcid.org/0000-0001-6742-4199

Fecha de recepción: 23-1-22
Fecha de aceptacion: 15-3-22

Resumen: *Este artículo propone un análisis sobre las desigualdades originadas o reforzadas por el uso de los modelos y programas de aprendizaje automático que encuentran una de sus principales justificaciones en el principio de precisión predictiva, así como sobre algunas respuestas que cabe esperar del sistema jurídico para abordarlas. Este tema ha sido tratado fundamentalmente bien a través de la tutela del derecho a la privacidad y la protección de datos, bien mediante la elaboración de códigos éticos. Aquí se adoptarán dos perspectivas distintas a las anteriores. En primer lugar, el enfoque del riesgo, marco en el que se ha emplazado recientemente la regulación europea de la inteligencia artificial y desde el que se examinarán las obligaciones vinculadas a las prácticas consideradas de alto riesgo. En segundo lugar, la perspectiva del Derecho antidiscriminatorio; en este sentido, se analizarán las virtualidades y los límites del este sector de los ordenamientos jurídicos. La finalidad de este artículo es mostrar que si el Derecho antidiscriminatorio ha de responder a las prácticas más controvertidas de los modelos algorítmicos, ha de superar la visión binaria de la teoría antidiscriminatoria y ser interpretada como una teoría jurídica con capacidad para transformar la realidad social.*

Abstract: *This article proposes an analysis of the inequalities that arise from or are reinforced using artificial intelligence models that find one of their main*

* Este trabajo se ha realizado en el marco del proyecto GVPROMETEO2018-156, del programa de investigación Prometeo de la Generalitat Valenciana.

justifications in the principle of predictive accuracy, and some of the replies that can be expected from the legal system to deal with them. The approach to this issue has mainly been through the right to privacy and data protection, on the one hand, or the development of ethical codes, on the other. Here I adopt two different perspectives. First, the risk approach offers a framework in which the European regulation of artificial intelligence has recently been situated from a regulatory point of view. It examines the obligations linked to practices considered high-risk. Second, the perspective of anti-discrimination law; the potential and limits of this area of Law. The aim of this article is to show that the capacity of anti-discrimination law to respond to the most controversial practices of algorithmic models, it must overcome the binary vision of anti-discrimination theory and be interpreted as a legal theory oriented to transform social reality.

Palabras clave: inteligencia artificial, derechos humanos, riesgo, discriminación estructural, justicia
Keywords: artificial intelligence, human rights, risk, structural discrimination, justice

1. INTRODUCCIÓN. INTELIGENCIA ARTIFICIAL E INSUFICIENTE VINCULACIÓN A LOS DERECHOS HUMANOS

La ubicuidad de la inteligencia artificial[1] es un hecho incuestionable que actualmente se proyecta sobre todas las áreas de conocimiento y sobre la conducta humana. Sus implicaciones en el ámbito de los derechos humanos son especialmente plásticas, dado que puede impactar de diversas formas en el ejercicio de un amplio conjunto de derechos como las libertades de expresión y reunión, el control de los datos personales, la satisfacción de los derechos sociales y de otros bienes básicos, el acceso a la justicia, la garantía de la

[1] La Propuesta de Reglamento de la UE sobre inteligencia artificial, de 21 de abril 2021 define la inteligencia artificial en estos términos: "El *software* que se desarrolla empleando una o varias de las técnicas y estrategias que figuran en el Anexo I y que puede, para un conjunto determinado de objetivos definidos por seres humanos, generar información de salida como contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa" (artículo 3.1). "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" (Anexo I).

seguridad o el derecho al trabajo, así como en determinados actos jurídicos propios de la esfera privada –piénsese, por ejemplo, en la concesión de un crédito o en la suscripción de un contrato de seguro–. Lo cierto es que la aplicación de las tecnologías digitales a la esfera de los derechos humanos puede utilizarse y, de hecho, se utiliza para "automatizar, predecir, identificar, vigilar, detectar, singularizar y castigar"[2]. No me refiero exactamente a estos usos de la inteligencia artificial cuando operan como instrumento o apoyo de una decisión, sino más bien cuando predeterminan la decisión final de los poderes públicos, delimitando el margen de discrecionalidad a partir de los postulados contenidos en la programación[3]. En este sentido, cabe hacer hincapié en el uso de la automatización con fines predictivos del comportamiento humano y su capacidad para originar prácticas injustas o controvertidas, como son las que crean o perpetúan situaciones de discriminación de personas y colectivos en situación de desigualdad –particularmente, de desigualdad estructural–, o bien las que justifican decisiones basadas en criterios irrelevantes, o las que comportan predicciones incorrectas. La contracara de las potencialidades de estos sistemas automatizados para articular nuevas formas de acceso, ejercicio y defensa de los derechos es la posibilidad –bien real y también novedosa– de su vulneración[4].

La aparente neutralidad del proceso que ha dado en llamarse transformación digital no debería soslayar la profundidad de los cambios cualitativos y estructurales que puede generar en los sistemas jurídicos. Como apunta Boix[5], constituye una exigencia mínima prestar atención a las modificaciones cualitativas que debe incorporar el Derecho "para preservar la equidad y la justicia social, y actuar como un muro de contención de las desigualdades

---

[2] P. ALSTON, "The Digital Welfare State. Report of the Special Rapporteur on extreme poverty and human rights", A/74/493, 11 de octubre de 2019, pár. 3.

[3] A. BOIX PALOP, "Los algoritmos son reglamentos: la necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones", *Revista de Derecho Público: Teoría y Método,* vol. 1, 2020, p. 237. DOI: 10.37417/RPD/

[4] Los avances que aporta la inteligencia artificial al conocimiento, la investigación y el bienestar humano son indudables, así como los previsibles logros en la defensa y ejercicio de los derechos. Sobre esta última faceta, véase S. BARONA *Algoritmización del Derecho y de la Justicia. De la Inteligencia Artificial a la Smart Justice,* Valencia, Tirant Lo Blanch, 2021.

[5] A. BOIX PALOP, "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", cit., p. 233.

que, sobre criterios eficientistas, podrían derivarse de la generalización de este modelo de toma de decisiones" [6].

En el análisis de la inteligencia artificial y sus efectos han prevalecido hasta hoy dos aproximaciones teóricas: por una parte, su abordaje a partir del derecho a la vida privada y la protección de los datos personales, y, por otra, su tratamiento desde el prisma ético –los códigos éticos[7]–. Sin duda, el estudio del impacto de la inteligencia artificial en los derechos humanos puede adoptar diversas perspectivas, de las que he seleccionado tres.

La primera perspectiva –que servirá de contexto general de esta reflexión– coincide en realidad con la preocupación relativa a la eventual evolución de la aplicación de la inteligencia artificial al ámbito de los derechos –que, cabe afirmar, se desarrolla al margen de ellos, de su lógica, de sus principios y de sus criterios interpretativos, generando una suerte de compartimentos estancos–. De un lado, porque la tecnología evoluciona de forma autónoma y no se acompasa a la evolución de los sistemas normativos y, de otro, porque los juristas se han aproximado a la transformación tecnológica sectorialmente, atendiendo a la afectación del derecho a la privacidad, a los regímenes de vigilancia o a la discriminación. Alston, que se ha ocupado de examinar en profundidad la evolución –a su juicio preocupante– del Estado social digital, advierte que el "riesgo de llegar a una distopía digital es real" [8] no solo porque la mutación que la inteligencia artificial está provocando en el Estado de bienestar evidencia una insuficiente vinculación a los derechos humanos, sino también porque la inteligencia artificial es, en la práctica, una zona "exenta de derechos humanos" [9]. Ello se percibe especialmente cuando las exigencias básicas que subyacen a los derechos humanos se convierten en una excepción. Así ocurre en ámbitos como la discriminación y la injusticia

[6] A. BOIX PALOP, "De McDonald's a Google: la ley ante la tercera revolución productiva", *Teoría & Derecho*, núm. 1, 2007, pp. 141-145.

[7] A título ilustrativo, GRUPO INDEPENDIENTE DE EXPERTOS DE ALTO NIVEL SOBRE IA: *Directrices Éticas para una IA fiable,* junio 2018. *The ethics of artificial intelligence: Issues and initiatives,* marzo de 2020. EUROPEAN COMMISSION FOR THE EFFICIENCY OF JUSTICE (CEPEJ) *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environme*nt, diciembre de 2018. Por otra parte, estas propuestas éticas muestran ciertas insuficiencias para regular la justicia algorítmica, especialmente ante riesgos de tipo sistémico, de ahí las propuestas orientadas a reforzar los elementos jurídicos. Cfr. J. BLACK y A. MURRAY, "Regulating AI and Machine Learning: Setting the Regulatory Agenda", *European Journal on Law and Technology,* núm. 10 vol. 3, 2019, p. 7.

[8] P. ALSTON "The Digital Welfare State", cit., pár. 2.

[9] Ibid., pár. 35.

de los resultados y las decisiones, la transparencia y rendición de cuentas –que incluso es considerada innecesaria en algunos casos–, la alteración de la responsabilidad o la afectación de la información personal[10].

La segunda perspectiva sitúa el desarrollo algorítmico en el marco de la noción de riesgo y la aplicación de su lógica al ordenamiento jurídico. La capacidad exponencial de procesar datos que provienen de multitud de fuentes ha convertido en una realidad la posibilidad de realizar análisis predictivos y automatizar decisiones –por ejemplo, sobre la elegibilidad de una persona para acceder a un programa social, la gestión sanitaria, o la verificación de la identidad–, circunstancia que da lugar a nuevas fuentes de discriminación con características propias que encuentran su acomodo en una concepción preventiva y securitaria del Derecho. Como señala San Martín[11], la perspectiva del riesgo en el campo del Derecho reenvía a la idea de incerteza e inseguridad y al mismo tiempo constituye una técnica de gestión, un modo de dirigir y resolver la incertidumbre. La administración del riesgo mediante la inteligencia artificial opera del mismo modo, es decir, como un conjunto de prácticas crecientemente complejas y automatizadas que dan lugar a la denominada *digital risk society*, pero también como el tratamiento jurídico de los modelos algorítmicos desde el enfoque del riesgo. De ahí el interés por comprender qué tipo de racionalidad impone la perspectiva del riesgo sobre la inteligencia artificial y, a su vez, qué efectos derivan de sus diseños y decisiones. Sin embargo, hay que plantear las consecuencias que tiene para los derechos fundamentales esta perspectiva, prioritariamente preventiva y menos centrada en la sanción de las infracciones normativas. Esta es la razón de que analicemos la dimensión predictiva de conductas de la inteligencia artificial a la luz del Derecho antidiscriminatorio: el objetivo es identificar y comprender qué tipo de respuestas puede ofrecernos esta orientación.

La tercera perspectiva es, pues, el Derecho antidiscriminatorio. Las publicaciones sobre esta temática son muy numerosas, especialmente en el mundo jurídico anglosajón. Como se ha apuntado arriba, en este trabajo me interesa fundamentalmente responder a la siguiente pregunta: ¿qué cabe esperar del Derecho antidiscriminatorio en este campo y en qué condiciones?

[10] J. COBBE, "Administrative law and the machines of government: judicial review of automated public-sector decision-making", *Legal Studies*, 39(4), 2019, pp. 4-5. Disponible en: *https://ssrn.com/abstract=3226913* or *http://dx.doi.org/10.2139/ssrn.3226913*

[11] D. SAN MARTÍN, *El concepto de riesgo en la racionalización del derecho punitivo. Razón jurídica y gestión del riesgo en la administración de la peligrosidad*, Tesis Doctoral, Universidad de la Rioja, 2021, p. 298.

Considero que se trata de un sector del Derecho particularmente fértil para visibilizar la discriminación que tiene su origen en la utilización de algoritmos. Previsiblemente, esta tarea será posible si los modelos algorítmicos cuentan con garantías, prohibiciones claras de discriminación (en los datos y en las clasificaciones) y si existen mecanismos que posibiliten el examen, el conocimiento y el control de las distintas formas de discriminación, así como la articulación de respuestas ante las mismas, dado que los perfiles que adopta de la inteligencia artificial en el ámbito del Derecho están estrechamente relacionados con las posibilidades de controlarla y de atribuir responsabilidades en un marco de principios claro. A este respecto, plantearé que el Derecho antidiscriminatorio tendrá un alcance limitado si asumimos que su aplicación se circunscribe a la resolución de un problema individual, puntual o accidental de discriminación cuyo escenario clave es la litigación y que su finalidad es meramente reparadora, mientras que sus potencialidades se ampliarán si consideramos que su finalidad estriba en la transformación de las estructuras sociales que generan las desigualdades y en la emancipación de los sujetos, es decir, si puede suministrar una visión prioritariamente centrada en las causas de la discriminación estructural y en el cuestionamiento de la lógica que subyace al uso de la inteligencia artificial.

2. EL RIESGO COMO PERSPECTIVA REGULATIVA

La perspectiva del riesgo puede ser concebida como una clave justificativa y explicativa que proyecta una racionalidad específica en los escenarios en los que se aplica y que ha cristalizado en el denominado "giro securitario preventivo" de los sistemas jurídico-políticos[12]. La introducción de la inteligencia artificial en este contexto hace referencia tanto a la idea de incerti-

[12] Sobre el giro securitario preventivo en el ámbito penal y criminológico, la bibliografía es extensísima. Respecto a la proyección de la perspectiva del riesgo y el desarrollo algorítmico en el Derecho penal, remito a L. MARTÍNEZ GARAY y F. MONTES SUAY, "El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias", *Indret,* núm. 2, 2018; y L. MARTÍNEZ GARAY, "Peligrosidad, algoritmos y *due process*: el caso *State vs. Loomis*", *Revista de Derecho Penal y Criminología*, núm. 20, 2018, pp. 485-502. Un análisis muy interesante de la sentencia *State vs. Loomis* puede leerse en J.I. SOLAR CAYÓN, "Inteligencia artificial en la justicia penal: los sistemas algorítmicos de evaluación de riesgos", *Dimensiones éticas y jurídicas de la inteligencia artificial en el Estado de Derecho,* Alcalá de Henares, Universidad de Alcalá, 2020, pp. 125-172. En relación con el enfoque del riesgo en el Derecho administrativo y, concretamente, en el Derecho de extranjería como paradigma, véase A. SOLANES, "Del Derecho penal al Derecho administrativo del enemigo: la legislación de ex-

dumbre y al riesgo algorítmico al que se exponen determinados bienes jurídicos básicos cuanto al modo de afrontar situaciones de inseguridad; en este segundo sentido, la inteligencia artificial se presenta como una herramienta de gestión de riesgos originariamente basada en el paradigma matemático, la probabilidad, la estadística y el cálculo cuyo objetivo es hacer frente a la incertidumbre mediante el despliegue de funciones anticipatorias (prevención, precaución, preparación) inspiradas en una racionalidad de carácter instrumental[13]. El nuevo escenario ha propiciado el desarrollo de modelos matemáticos capaces de realizar operaciones tales como procesar a los sujetos, reconocer patrones, predecir conductas y adoptar decisiones a través de la estadística computacional. Como se ha señalado reiteradamente, estos modelos han desplazado a la probabilidad clásica basada en el principio de causalidad y la han sustituido por un tipo de razonamiento que recurre a correlaciones múltiples y no evidentes y que, por tanto, provoca un aumento de la opacidad, uno de los rasgos característicos de su funcionamiento. Esta forma de proceder, escribe San Martin[14], incrementa la eficiencia, pero tiene un precio: la eventual introducción de sesgos y de criterios prohibidos por el Derecho antidiscriminatorio en la motivación jurídica.

Pues bien, desde el punto de vista jurídico[15] constituye un objetivo pertinente abordar los rasgos característicos de la aplicación de la inteligencia artificial al Derecho –entre ellos, la opacidad, la complejidad técnica, los sesgos, cierta imprevisibilidad y el comportamiento parcialmente autónomo de algunos sistemas de inteligencia artificial– con el objetivo de analizar en qué medida está garantizada su compatibilidad con los derechos fundamentales y hasta qué punto facilitan la aplicación de las normas jurídicas. La perspectiva consiste en "situarse ante las diversas fuentes de riesgo, mediante un enfoque basado también en el riesgo"[16].

---

tranjería como ejemplo", en *Estudios jurídicos en memoria de la profesora Elena Górriz*, Valencia, Tirant Lo Blanch, 2020, pp. 801-819.

[13] D. SAN MARTÍN, *El concepto de riesgo en la racionalización del derecho punitivo,* cit., pp. 222-223.

[14] *Ibid*., pp. 264 y 307.

[15] Conclusiones del Consejo de la Unión Europea sobre la Carta de los Derechos Fundamentales en el contexto de la inteligencia artificial y el cambio digital, 11481/20, 21 de octubre de 2020.

[16] Así se expresa la Propuesta de Reglamento Europeo sobre Inteligencia artificial en su párrafo 3.5.

Este parece ser el marco en el que se ubica la Propuesta de Reglamento de la Unión Europea orientada a armonizar la legislación en materia de inteligencia artificial[17], instrumento que plantea una visión más amplia que la que hasta el momento ofrecía la normativa de protección de datos en el espacio europeo. Al menos desde el punto de vista de su enfoque, cabe sostener que el proyecto está en consonancia con las advertencias de Yeung sobre las implicaciones jurídicas, sociales y democráticas de la inteligencia artificial y sobre el hecho de que en su aplicación están en juego cuestiones que trascienden el derecho a la privacidad, el control de los datos personales e incluso los principios de transparencia y de responsabilidad[18].

La exposición de motivos de la propuesta de Reglamento afirma que este se orienta "a buscar un enfoque equilibrado" entre los intereses del mercado europeo y los derechos de los ciudadanos y que su ámbito de aplicación se extiende a los ámbitos público y privado. En el texto, la inteligencia artificial es caracterizada como una oportunidad a la que no se puede renunciar por los beneficios asociados a su operatividad y eficiencia, pero que puede vulnerar o poner en riesgo derechos fundamentales y bienes jurídicos protegidos, incluso especialmente protegidos.

El enfoque equilibrado por el que aboga texto –un equilibrio que, conjeturo, no estará exento de dificultades– se sustenta en una "regulación proporcional"[19] que trata de conciliar, por una parte, el entramado normativo existente sobre las garantías de los derechos fundamentales, la legislación sobre protección de datos y sobre propiedad intelectual e industrial, y, por otra, un enfoque basado en el riesgo que modula el establecimiento de obligaciones en función del grado de amenaza para los derechos y la seguridad.

---

[17] UE, Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión, COM (2021) 206 final, 2021/01106 (COD), 21/4/2021.

[18] K. YEUNG, "Algorithmic Regulation: A Critical Interrogation", *Regulation & Governance*, núm. 12 vol. 4, 2018, pp. 518-159. DOI:10.1111/rego.12158

[19] Así se reconoce en la exposición de motivos (punto 2.3). Por otra parte, la idea de equilibrio se plasma en distintos aspectos, uno de los cuales es la exigencia de transparencia. La propuesta de Reglamento no exige una transparencia total, sino un grado de transparencia compatible con el cumplimiento de las obligaciones legales del usuario y del proveedor. El equilibrio estriba en el hecho de que el texto establece el deber del proveedor de mostrar cómo funciona la aplicación, incluida su "lógica general" –asi como los presupuestos de partida o una descripción de los datos que se hayan utilizado para su creación–, pero no le exige una absoluta transparencia sobre el *sofware* utilizado.

Entre las técnicas o instrumentos de intervención previstas en la propuesta de Reglamento cabe destacar la clasificación de las conductas de acuerdo con los niveles de riesgo, la prohibición de determinadas prácticas que se consideran inaceptables a fin de evitar los riesgos través de la aplicación del principio de precaución, el tratamiento de los sistemas de alto riesgo mediante obligaciones impuestas a los operadores y el establecimiento de normas sobre transparencia y sobre control y vigilancia del mercado (art. 2). Asimismo, prevé formas de control preventivo –entre ellas, un régimen de autorizaciones– o de control *ex post* –a través de la responsabilidad civil y/o penal–, un sistema de auditoria y de inspección que, en general, se activa a instancia de los perjudicados. Por lo que respecta los sistemas que no son considerados de alto riesgo, el texto de la propuesta establece obligaciones limitadas en materia de transparencia.

Como he señalado, la propuesta de Reglamento sigue un enfoque basado en los riesgos que distingue entre aquellos usos de la inteligencia artificial que generan i) un riesgo inaceptable, ii) un riesgo alto, y iii) un riesgo bajo o mínimo que deriva no tanto de la implementación de medidas cuantitativas, sino más bien cualitativas. Así, el Título II, que hace referencia a todos los sistemas algorítmicos, enumera una serie de prácticas prohibidas cuyo uso se considera inaceptable debido a su potencial para ocasionar daños o poner en peligro bienes claramente protegidos a través de los derechos fundamentales o la seguridad.

Entre estas prácticas cabe mencionar las siguientes: la utilización de sistemas que recurren a alguna técnica subliminal que modifique de manera sustancial el comportamiento de las personas provocando daños a terceros o a ella misma; el aprovechamiento de la vulnerabilidad de una persona o grupo para alterar de manera sustancial su comportamietno de forma que origine un perjuicio a ella misma o a otras personas; o la identificación biomética. Dejando ahora al margen la calidad técnica de este proyecto de regulación, la imprecisión de las conductas que resultarán efectivamente prohibidas y las dudas sobre cómo pueden ser evaluadas desde el punto de vista del principio de precaución, centraré la atención en uno de los apartados de este precepto.

El supuesto contemplado en el artículo 5, apartado (c) de la propuesta de Reglamento prohíbe "la introducción en el mercado, la puesta en servicio o la utilización de sistemas de IA por parte de las autoridades públicas o en su representación con el fin de evaluar o clasificar la fiabilidad de personas

físicas durante un período determinado de tiempo atendiendo a su conducta social o a características personales o de su personalidad conocidas o predichas, de forma que la clasificación social resultante provoque una o varias de las situaciones siguientes: (i) un trato perjudicial o desfavorable hacia determinadas personas físicas o colectivos enteros en contextos sociales que no guarden relación con los contextos donde se generaron o recabaron los datos originalmente; (ii) un trato perjudicial o desfavorable hacia determinadas personas físicas o colectivos enteros que es injustificado o desproporcionado con respecto a su comportamiento social o la gravedad de este".

Por su parte, el Anexo III de la propuesta de Reglamento presenta un listado de sistemas de "alto riesgo" que concreta los ámbitos en los que pueden realizarse estas conductas: 1. Identificación biométrica y categorización de personas físicas. 2. Gestión y funcionamento de instraestructuras esenciales. 3. Educación y formación profesional. 4. Empleo, gestión de trabajadores. 5. Acceso y disfrute de servicios públicos y privados esenciales y sus beneficios. 6. Asuntos relacionados con la aplicaciónd de la ley. 7. gestión de la migración, asilo y control fronterizo; y 8. Administración de justicia y procesos democráticos[20].

Es posible que en la mayoría de estas esferas se produzcan usos de los sistemas de inteligencia artificial que clasifiquen a las personas de acuerdo con los criterios prohibidos por el arriba reproducido apartado (c). Por ejemplo, que se lleven a cabo clasificaciones sociales sobre la base de las predicciones elaboradas a partir del comportamiento de un grupo de población general o mediante correlaciones con un conjunto de datos que aparentemente no tienen relación con la decisión a adoptar –el acceso a un empleo, el aumento o no de una prima de seguro, entre ellas–. El resultado es que las clasificaciones del riesgo y de las personas pueden reforzar o agravar las desigualdades y la discriminación que padecen ciertos colectivos[21]; tal es el caso, por ejemplo, de los beneficiarios de un programa social que se ven compelidos a renunciar a su derecho a la privacidad para mantener su derecho a percibir la prestación.

[20] J.I. SOLAR CAYÓN, "Retos de la deontología de la abogacía en la era de la inteligencia artificial jurídica", *Derechos y Libertades,* núm. 45, 2021, pp. 123-161.

[21] V. EUBANKS, *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*, Londres, St. Martin's Press, 2019; y L DENCIK, A. HINTZ, J. REDDEN, E TRERÉ, "Exploring Data Justice: Conceptions, Applications and Directions, Information", *Communication & Society*, núm. 22 vol. 7, 2019, 873-881.DOI: 10.1080/1369118X.2019.1606268

Ahora bien, como señala Huergo[22] en relación con la características del proyecto de regulación, al menos los tres primeros supuestos previstos constituyen, en realidad, versiones extremas de algunos usos habituales de los modelos algorítmicos que aparecen descritos en términos tan exagerados y con tal acumulación de adjetivos que su aplicación se antoja difícil y su utilidad limitada. Por ello, cabe conjeturar que, en la medida en que cada uno de los supuestos da lugar a un parámetro normativo complejo e indeterminado, no resultará fácil establecer el grado de inobservancia de las prohibiciones y la exigencia de responsabilidad.

En su exposición de motivos, la propuesta de Reglamento señala expresamente el objetivo de establecer obligaciones proporcionadas a cuantos intervienen en la "cadena de valor" con la doble finalidad de promover la protección de los derechos fundamentales y conseguir que la programación algorítmica sea fiable[23]. Estas obligaciones se dirigen tanto a los poderes públicos como a los agentes privados. Sin embargo, la determinación de los supuestos en los que los poderes públicos han de establecer obligaciones y reglas que deben cumplir los actores privados adolece de una notable imprecisión.

No parece que esta propuesta avance hasta el punto de considerar el modelo algorítmico como una norma jurídica, dado que hace referencia a ella como una "documentación técnica" que debe acompañar a los sistemas de alto riesgo (art. 11 y Anexo III). Es conocida la propuesta de Boix orientada a tratar el sistema o modelo algorítmico como una norma jurídica[24]. Concretamente, el autor se refiere a una norma de tipo reglamentario como la que prevé el ordenamiento jurídico español. Boix parte de la idea de que la utilización predictiva y decisoria (no meramente instrumental) de los algoritmos predetermina la decisión final del poder público y limita el ámbito de discreción o de capacidad de determinación de quienes deben aplicarlos a partir de los postulados

[22] A. HUERGO, "El proyecto de reglamento sobre la inteligencia artificial", *El Almacén de Derecho*, 17 de abril de 2021, p. 3. Disponible en*: https://almacendederecho.org/el-proyecto-de-reglamento-sobre-la-inteligencia-artificial/*. Acceso: 20/09/2021.

[23] Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial, cit.*,* p. 12.

[24] "Deben ser tratados como normas reglamentarias si la tecnología es empleada para adoptar decisiones administrativas o son apoyo esencial de las mimas, bien a la hora de evaluar las circunstancias concurrentes –decantación del supuesto de hecho o bien a la hora de aportar evaluaciones o elementos de juicios sobre cual pueda ser la mejor o la más apropiada medida a adoptar (consecuencia jurídica)" (A. BOIX PALOP, "Los algoritmos son reglamentos…", cit., p. 237).

contenidos en la programación. Se trata de los elementos esenciales de una norma reglamentaria. En estos casos, "la programación algorítmica ha de ser considerada norma jurídica (reglamentaria) pues tal es su función material, estrictamente equivalente a la programación jurídica en la que en muchos casos se integra o a la que, en muchos otros casos, poco a poco aspira a sustituir". En realidad, como sugiere Solar Cayón, la transparencia –y, por tanto, la publicidad– de la norma jurídica exige que su redacción refleje tanto el diseño del algoritmo como los datos de su entrenamiento[25].

Si prevalece la finalidad de favorecer el uso de algoritmos y programas tanto por parte de las autoridades como en la esfera privada, a la hora de regular el funcionamiento de la inteligencia artificial en los sistemas jurídicos lo fundamental es reforzar las garantías de los derechos fundamentales y no debilitarlas. Algunos aspectos de la propuesta de Reglamento inducen a pensar que el principio de equilibrio que parecía inspirar toda la regulación no fomenta, en realidad, el equilibrio prometido, dado que la perspectiva adoptada por el texto es fundamentalmente la de gestión de riesgos y que la robustez de las garantías de los derechos es insuficiente.

A este respecto, Grace y Bamford abogan por el establecimiento de una regulación sobre la inteligencia artificial a través de un programa legislativo sectorial capaz de construir (y robustecer) una "cultura jurídica algorítmica"[26]. Partiendo de determinados principios de justicia, consideran que, del mismo modo que no se aprobó una única legislación para regular la revolución industrial, tampoco la regulación de la toma de decisiones algorítmicas puede plasmarse en un solo texto legal. Dado que en los distintos sectores sociales los riesgos son diferentes y están en juego diferentes normas y valores, ambos autores sugieren la pertinencia de articular una legislación parcelada o sectorial que establezca obligaciones y garantías precisas para reforzar el entramado jurídico existente.

Seguramente tiene razón Alston cuando señala que nos encontramos frente a procesos que no tienen precedentes y esto hace que no siempre contemos con conceptos precisos para comprenderlos adecuadamente o no cuestionarlos[27].

---

[25] J.I. SOLAR CAYÓN, "Inteligencia artificial en la justicia penal…", cit., p. 157.

[26] F.J., ZUIDERVEEN BORGESIUS, "Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence", *The International Journal of Human Rights*, núm. 24 vol. 10, 2020, pp. 28-29.

[27] P. ALSTON "The Digital Welfare State", cit., pár. 69.

Si examinamos la propuesta de Reglamento desde el punto de vista de las garantías y las obligaciones previstas en el texto, pueden identificarse tres cuestiones que reclaman una respuesta: (a) En relación con las obligaciones, planteo si es suficiente una obligación genérica de diligencia debida. (b) Con respecto a la transparencia y la comprensibilidad de los algoritmos, cabe preguntarse por la opacidad de los sistemas de aprendizaje y el equilibrio entre transparencia y precisión y (c) La tercera cuestión atañe a la pretensión de imparcialidad y los sesgos.

(a) Las obligaciones orientadas al cumplimiento del objetivo de prevenir daños a bienes y derechos fundamentales cristalizan, en realidad, en la obligación de diligencia debida en el marco de los límites normativos ya existentes[28]. Del conjunto de obligaciones y garantías concretas en relación con las conductas de alto riesgo y las conductas prohibidas –así, la declaración responsable o de conformidad (Anexo V) y el cumplimiento de una serie de requisitos (entre ellos, la presentación de una documentación técnica sobre la aplicación que ha de cumplir los criterios de solidez, exactitud y seguridad (Anexo IV)[29], la propuesta obliga a garantizar la dirección humana y tener un grado de transparencia suficiente. Huergo[30] sostiene que, a diferencia de otros ámbitos de riesgo tradicionales, en los que las normas establecen medidas concretas de seguridad, aquí se establecen unos requisitos muy generales que, en realidad, operan como una declaración responsable o como una forma de cumplimiento normativo. En los supuestos de los sistemas que no son de alto riesgo, tan solo se imponen obligaciones limitadas en materia de transparencia –por ejemplo, la presentación de información para comunicar el uso de un sistema de IA cuando este interactúe con humanos–. Por otra parte, el mismo autor advierte que, aun cuando resulte posible un control *ex post* a través de la exigencia de responsabilidad civil o penal, es necesario valorar también en qué medida la responsabilidad quedaría excluida al aplicarse estas medidas, que serían consideradas una actuación diligente dirigida a minimizar los riesgos[31].

Queda lejos, por tanto, el establecimiento de una obligación directa que derivaría de la tendencial asimilación del algoritmo a una norma jurídica.

---

[28] A. HUERGO, "El Proyecto de Reglamento sobre la inteligencia artificial…", cit., p. 7.

[29] El Anexo IV de la Propuesta de Reglamento detalla la documentación técnica que debe acompañar a los sistemas de inteligencia artificial.

[30] A. HUERGO, "El proyecto de reglamento sobre la inteligencia artificial…", cit., p. 7.

[31] Ibid., p. 10.

Me refiero a la obligación dar cumplimiento al principio de publicidad del algoritmo o código fuente, tal y como propone Boix[32], una exigencia que la popuesta de Reglamento debería haber contemplado para establecer garantías concretas a los ciudadanos frente a las actuaciones de los poderes públicos (y privados) que pudieran afectar a los derechos fundamentales, particularmente en materia de igualdad y prohibición de discriminación.

(b) En relación con la segunda cuestión planteada, la propuesta de Reglamento declara que la documentación, la trazabilidad, la transparencia, la precisión y la solidez son exigencias para reducir los riesgos de la inteligencia artificial en relación con los derechos fundamentales y la seguridad que no están cubiertos por otros marcos jurídicos existentes.

Sin embargo, la inclusión del uso de la inteligencia artificial en el marco interpretativo del riesgo no zanja la cuestión de la racionalidad decisoria y las garantías de la misma. De acuerdo con la doctrina dominante, el principio de precaución procede cuando se parte de un déficit de información o de capacidad técnica o cuando existen obstáculos para poder valorar las consecuencias de una decisión. Sin embargo, la utilización de programas y modelos algorítmicos ha transformado la aplicación del principio de precaución. Como explica San Martín[33], en la perspectiva tradicional del riesgo, el principio opera si las consecuencias son calculables y si existe acuerdo sobre su clasificación, de forma que, cuando las consecuencias son calculables pero no hay acuerdo sobre ellas, nos movemos en terreno de la ambigüedad; en cambio, si hay acuerdo sobre las consecuencias, pero el cálculo no es posible, nos situamos en el ámbito de la incertidumbre. Aquí es donde se ha producido el cambio. En la forma de operar de la inteligencia artificial, "el principio de causalidad se diluye ante puras correlaciones, reconocimiento de patrones y la multiplicación exponencial de las interacciones entre datos, incorporados en volúmenes masivos con modelos muy abstractos que se entrenan a si mismos", dando lugar a sistemas muy complejos que generan opacidad –o a los denominados algoritmos de caja negra, que también derivan de la protección de intereses contrapuestos–. Esta forma de actuar, especialmente en procesos que afectan a los derechos de las personas, dificulta la observancia de las exigencias de transparencia y publicidad de los criterios normativos que han servido de base para la toma de decisiones.

---

[32] A. BOIX PALOP, "Los algoritmos son reglamentos…", cit., pp. 262-265.

[33] D. SAN MARTÍN, *El concepto de riesgo en la racionalización del derecho punitivo,* cit., p. 268.

Ante esta situación, se proponen distintas iniciativas. Alston[34], por ejemplo, señala que "para garantizar que se tengan en cuenta las consideraciones relativas a los derechos humanos", es necesario velar por que las prácticas en que se basa la creación, la auditoría y el mantenimiento de los datos se sometan a un escrutinio muy intenso. Por su parte, San Martín[35] subraya que, aunque importa no abandonar la comprensión de la mecánica interna de los algoritmos, es preciso centrarse en la valoración de las consecuencias jurídicas de su opacidad y en las posibilidades de preservar las garantías de los destinatarios de las decisiones adoptadas a través de estos dispositivos. Esto es especialmente relevante en aquellos casos que en los que hay que identificar los sesgos derivados del proceso de entrenamiento del algoritmo y que pueden resultar difíciles de detectar en fase de diseño[36]. Más abajo trataré de argumentar que estas dos opciones son relevantes y que han de encauzarse hacia el fortalecimiento de los principios de justicia, de ahí la llamada de atención sobre la necesidad insoslayable de que los poderes públicos sean extremadamente cuidadosos en el manejo de la información automatizada y la valoración de los efectos de las decisiones basadas en ella[37].

La tercera cuestión (c) será abordada en el apartado siguiente. Mediante su análisis trataré de analizar un tema que permanece abierto, a saber, cómo puede mantenerse el equilibrio entre derechos, garantías y eficiencia en el uso de la inteligencia artificial por parte de los poderes públicos. La propuesta de Reglamento europeo establece ciertas prohibiciones que son algo indeterminadas; entre las conductas de alto riesgo, se encuentran las que pueden dar lugar a clasificaciones discriminatorias. En este sentido, propondré una reflexión sobre las respuestas que puede suministrar el Derecho antidiscriminatorio al respecto.

Pensemos en el conocido caso *Netherlands Committee of Jurists for Human Rights vs. State of the Netherlands* (2020)[38]. En los Países Bajos fue anulado por un tribunal el uso de un sistema automatizado llamado SyRI que se utilizaba para detectar preventivamente la probabilidad de que las personas incu-

---

[34] P. ALSTON, "The Digital Welfare State…", cit., pár. 119.

[35] D. SAN MARTÍN, *El concepto de riesgo*…, cit., pp. 305 y 311.

[36] J.I. SOLAR CAYÓN, "Inteligencia artificial en la justicia penal…", cit., p. 159.

[37] N. JANSEN, *On digital technology, social protection and human rights. A report from the Netherlands*, 31/05/2019. Disponible en: *https://www.ohchr.org/Documents/Issues/Poverty/DigitalTechnology/ITCU.pdf*. Acceso: 15/09/2021.

[38] *Netherlands Committee of Jurists for Human Rights vs. State of the Netherlands* (2020) ECLI: NL: RBDHA: 2020, p. 865.

rrieran en fraude para adquirir la condición de beneficiarios de prestaciones sociales –en el asunto que analizamos, subsidios por hijos menores a su cargo–. Se trataba de un sistema de detección basado en el riesgo que fue implantando fundamentalmente en zonas económicamente desfavorecidas y caracterizadas por sus altas tasas de población inmigrante[39]. Las autoridades administrativas del país acusaron erróneamente de fraude a muchas familias perceptoras de la prestación. En la utilización del modelo se hizo uso de datos como el origen de las familias y el sistema apuntó a los hogares de ascendencia marroquí y turca. El resultado fue, para muchas familias, la ruina económica, dado que se les obligó a devolver las prestaciones que habían recibido y fueron denegadas las nuevas solicitudes de prestación que presentaron, ello al margen del estigma social que sufrieron por un fraude que no habían cometido[40]. El examen judicial de la aplicación del sistema automatizado SyRi es de gran importancia, dado que el caso dio lugar a la primera sentencia de un tribunal europeo en la que ha sido sometido a examen un sistema algorítmico de evaluación del riesgo. La sentencia concluyó que se había producido (a) la afectación del derecho humano a la vida privada a través de la violación del articulo 8 del Convenio Europeo de Derechos Humanos[41]; y (b) la vulneración de la exigencia de transparencia. En respuesta a las alegaciones que apuntaban a la opacidad del algoritmo, el Gobierno neerlandés argumentó que el conocimiento de este sistema tendría como consecuencia que las personas ajustaran su conducta a sus parámetros, argumento que fue desestimado por el tribunal[42]. Hay que subrayar, en todo

[39] N. JANSEN, "Tackling the Human Rights Impacts of the "Digital Welfare State"", *Digital Freedom Fund*, 10/02/2020. Disponible en: *https://digitalfreedomfund.org/tackling-the-human-rights-impacts-of-the-digital-welfare-state/*. Acceso 20/7/2021.

[40] N. JANSEN, "On digital technology, social protection and human rights…", cit., p. 6. El Defensor del Pueblo holandés presentó un informe al respecto en 2017 y también el Parlamento Holandés debatió un informe interno en marzo de 2019 sobre este programa, que se aplicó en 2014.

[41] Alston analiza este sistema y subraya que constituye un ejemplo del riesgo de singularizar y acosar deliberadamente a los pobres mediante el uso de las nuevas tecnologías en el marco del Estado del bienestar. A menudo, continua el autor, en el Estado del bienestar el fraude es el resultado de la confusión, la complejidad y la incapacidad para corregir los errores provocados por las nuevas tecnologías, que propician el fenómeno que ha sido denominado la "muerte de la amnesia": la capacidad de recopilar información y almacenarla digitalmente durante un período indefinido posibilita la utilización *sine die* de gran cantidad de datos contra una persona (P. ALSTON, "The Digital welfare state…", cit., pár. 103).

[42] J. GRACE Y R. BAMFORD, ""AI Theory of Justice": using rawlsian approaches to legislate better on Machine Learning in Governmet", *Amicus Curiae*, num. 1 vol. 3, 2020, pp. 355-356.

caso, que el supuesto no fue abordado como un supuesto de discriminación estructural (institucional). Probablemente, el recurso a este tercer examen habría supuesto mayores avances en el control de las decisiones algorítmicas.

3. EL DERECHO ANTIDISCRIMINATORIO COMO RESPUESTA Y SUS LÍMITES

Como he señalado al comienzo de este trabajo, el uso de sistemas automatizados predictivos del comportamiento humano para la toma de decisiones puede contribuir a reforzar los procesos de discriminación existentes, pero también a generar nuevos supuestos de discriminación[43]. De acuerdo con lo que se ha expuesto hasta aquí, determinados usos de estos sistemas podrían constituir una conducta prohibida o una conducta de alto riesgo para los derechos. Me parece interesante señalar preliminarmente que los distintos enfoques sobre el desarrollo algorítmico – el ético, el técnico e incluso el del riesgo – se caracterizan por dos tendencias: por una parte, se centran en encontrar soluciones circunscritas a la minimización de daños; por otra, asumen la premisa de que los perjuicios que generan son individuales. Sin embargo, los perjuicios son sociales y estructurales: en este punto se encuentra la clave para comprender la lógica subyacente, las estructuras de privilegios y los resultados injustos de aquellas soluciones[44]. Veamos, pues, si el Derecho antidiscriminatorio tiene capacidad para ofrecer alguna respuesta en este sentido[45].

---

[43] Sobre ello advirtieron tempranamente C. O'NEIL, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Nueva York, Ramdom House, 2017; y V. EUBANKS, *Automating Equality: How High-Tech Tools Profile, Police and Punish the Poor*, cit.

[44] U. ANEJA, "La gobernanza de la inteligencia artificial: de solucionar los problemas a diagnosticarlos", *CIDOB*, julio 2021, pp. 28-35. Disponible en:

https://www.cidob.org/articulos/anuario_internacional_cidob/2021/la_gobernanza_de_la_inteligencia_artificial_de_solucionar_los_problemas_a_diagnosticarlos. Acceso: 10/09/2021.

[45] Alba Soriano ha sido una de las primeras autoras que en España ha abordado en profundidad el tratamiento jurídico de la discriminación algorítmica y la respuesta a la misma desde el Derecho antidiscriminatorio. En este apartado prestaré especial atención a sus dos principales trabajos sobre el tema: A. SORIANO, "Decisiones automatizadas y discriminación: aproximación y propuestas generales", *Revista General de Derecho Administrativo*, num. 56, 2021, pp. 1-45; y "Decisiones automatizadas: problemas y soluciones jurídicas. Más allá de la protección de datos", *Revista* de *Derecho Público: Teoría y Método*, num. 1 vol. 3, 2021, pp. 58-127. DOI:10.37417/RPD/

En términos generales, los resultados discriminatorios se producen porque existen sesgos en los datos y/o en la clasificación[46]. Hay sesgos en los datos cuando se utiliza directamente un motivo prohibido o cuando se elije una variable que correlaciona en mayor medida que otras con uno de los motivos prohibidos o un dato basado en rasgos irrelevantes pero que pueden tener los mismos efectos que los rasgos prohibidos[47]. Asimismo, cabe hablar de sesgos en los datos cuando se reproducen los prejuicios mostrados en ejemplos durante el entrenamiento del algoritmo[48] y cuando se extraen consecuencias negativas sobre los grupos protegidos a partir de muestras no representativas –bien porque se elige un conjunto demasiado limitado de características que no incluye a todos los grupos poblacionales o a todos los que pertenecen a un grupo (infrainclusión), bien porque los sobrerrepresentan[49]–.

Por otra parte, hay sesgos de clasificación. Hemos visto que los sistemas automatizados pueden emplearse con el fin de clasificar a los seres humanos y hacer predicciones sobre su comportamiento futuro. Ocurre que muchas de las características seleccionadas para llevar a cabo este cometido tienden a mimetizar o reproducir la forma en la que funcionan de hecho las sociedades y las estructuras que sitúan a las personas y los grupos en posiciones y estatus desiguales, asumiendo como criterios neutrales aquellos que, en realidad, constituyen un reflejo de estas posiciones de desigualdad, de subordinación y desventaja[50]. De este modo, los algoritmos

[46] A. SORIANO, "Decisiones automatizadas: problemas y soluciones jurídicas…", cit., pp. 89-93.

[47] S. BAROCAS y A. D. SELBST, "Big Data's Disparate impact", *California Law Review,* vol. 104, 2016, pp. 678-680. DOI: *http://dx.doi.org/10.15779/Z38BG31*

[48] S.M., WEST, M. WHITTAKER, M. y K. CRAWFORD, *Discriminating Systems: Gender, Race and Power in AI.* AI Now Institute. 2019. Disponible en: *https://ainowinstitute.org/discriminatingsystems.pdf.* Acceso: 20/07/2021. Las autoras hacen referencia a los sesgos de los propios programadores. No se trata de una cuestión trivial; de hecho, se considera que una vía privilegiada para evitar los sesgos de los algoritmos consiste en asegurar la diversidad de género, de origen étnico y socioeconómica entre las personas que trabajan como programadores. Por su parte, Solar hace referencia a un informe enmarcado en la misma línea que fue elaborado por la Cámara de los Lores, *AI in the UK: ready, ¿willing and able?* Disponible en: *https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf* (J. I. SOLAR CAYÓN, "Inteligencia artificial en la justicia penal…", cit., p. 165)

[49] Sobre los procesos de generalización en el Derecho véase F. SCHAUER, *Las reglas del juego. Un examen filosófico de la toma de decisiones basada en reglas en el derecho y en la vida cotidiana*, trad. C. Oronesu y J.L. Rodríguez, Madrid, Marcial Pons, 2004, pp. 75-96.

[50] Al respecto, se ha señalado que "los algoritmos son tan buenos como los datos de los que se nutren" (S. BAROCAS y A.D. SELBST, "Big Data's Disparate impact", cit., p. 730). La

replican y perpetúan las estructuras de discriminación y los estereotipos que subyacen a las mismas y tienen efectos negativos para los derechos de los grupos discriminados y para sus posibilidades de participación en la sociedad, dado que también cada vez es más sencillo identificar o singularizar a personas, grupos y segmentos sociales sobre los que recae la discriminación. Por tanto, la discriminación potencial que anida en la utilización de la inteligencia artificial puede perjudicar a personas con características protegidas y también puede comportar desventajas para determinados sujetos sobre la base de la selección de rasgos que correlacionan con los motivos protegidos, dando lugar a nuevas discriminaciones. Podrían, por ejemplo, reforzar las desigualdades socioeconómicas a partir de la aplicación de criterios aparentemente irrelevantes o conducir a predicciones incorrectas sobre la conducta individual[51].

De acuerdo con numerosas recomendaciones institucionales[52], la constatación de estas posibles fuentes de discriminación aconseja que la utilización de la inteligencia artificial en la adopción de decisiones se guíe por dos criterios. Por un lado, la evaluación cuidadosa del diseño y el desarrollo de modelos algorítmicos a fin de identificar qué normas de derechos humanos y antidiscriminatorias pueden verse afectadas a consecuencia de la calidad de los datos que se introducen y extraen. Por otro, la valoración de la procedencia y las posibles deficiencias del conjunto de datos, de la posibilidad de su uso inadecuado o descontextualizado, de las externalidades negativas resultantes de estas deficiencias, así como de los entornos en los que se utilizará o podría utilizarse el conjunto de datos.

Por lo que respecta a las características de la "discriminación algorítmica", de lo expuesto hasta aquí puede inferirse claramente que tiene, al menos, tres propiedades básicas que es necesario exponer sintéticamente para poder

---

cuestión no solo es que plasmen los prejuicios de los responsables y los programadores –tal vez no sea gratuito recordar algo tan trivial como que en este ámbito la diversidad de género y etnia es muy limitada–, sino sobre todo que reflejen los prejuicios generalizados que persisten en la sociedad a través de la constatación de regularidades que constituyen patrones preexistentes de exclusión, desigualdad y segregación social.

[51] J. GERARDS y F. ZUIDERVEEN, "Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence", *Colorado Technology Law Journal*, *preprint*, 2 de noviembre de 2020, pp. 14-16 y 62 y ss.

[52] Por ejemplo, CONSEJO DE EUROPA, Recommendation CM/Rec (2020)1 of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems 8 April Strasbourg: Committee of Ministers.

valorar la tutela antidiscriminatoria[53]: (a) la *invisibilidad*, que hace referencia a la opacidad de los datos y algoritmos así como a la indeterminación de la responsabilidad, (b) la *intensidad*, que denota la capacidad de multiplicar los efectos discriminatorios o reproducir las desigualdades, y (c) la *complejidad técnica*: los algoritmos pueden ser discriminatorios, pero también pueden serlo las decisiones adoptadas a partir de ellos[54].

El Derecho antidiscriminatorio constituye un subsistema jurídico que, como se ha dicho, puede ser fértil ante determinadas implicaciones del desarrollo de la inteligencia artificial si es capaz de superar ciertos límites que están presentes en su propia evolución y que la discriminación algorítmica ha intensificado. Por ello, resulta pertinente recordar sus objetivos o finalidades, así como la distinción entre discriminación directa e indirecta.

Como es sabido, en el ámbito del Derecho antidiscriminatorio europeo la dicotomía discriminación directa/discriminación indirecta es una clasificación estándar y asentada legislativa y jurisprudencialmente, aun cuando persisten debates abiertos sobre la misma. A mi juicio, resulta pertinente cuestionar la tesis de que esta dicotomía configura una clasificación exhaustiva y excluyente. La legislación y la jurisprudencia europeas en la materia son deudoras de la tradición jurídica estadounidense. Sin embargo, como advierte Barrère, la identificación entre discriminación directa e indirecta con el *disparate treatment* y el *disparate impact* de la jurisprudencia norteamericana "resultó excesivamente precipitada"[55]. Ello es así porque, en realidad, la reconstrucción de la jurisprudencia del Tribunal Supremo estadounidense –prosigue Barrère– identifica tres ejes de discriminación, no dos, y en ellos se dan criterios heterogéneos y no simétricos que, por tanto, no pueden dar lugar a una clasificación dicotómica. En todo caso, la clasificación no responde a la diferencia entre el *disparate treatment* –que se identifica por la intencionalidad del trato– y el *disparate impact* –que se define por la neutralidad (o la indiferenciación) del trato y el efecto grupal diferenciado–.

Sea como fuere, la clasificación discriminación directa/indirecta se integró en el Derecho europeo y ha cristalizado en un esquema binario que, de

[53] C. SÁEZ, "El algoritmo como protagonista de la relación laboral. un análisis desde la perspectiva de la prohibición de discriminación", *Temas laborales,* núm. 155, 2020, p. 45.

[54] A. SORIANO, "Decisiones automatizadas y discriminación…", cit., pp. 11-14.

[55] Sobre el examen de esta evolución en la cultura jurídica estadounidense y en la europea, remito a M. A. BARRÈRE, "El principio de igualdad de trato vinculado a la discriminación directa e indirecta", *Feminismo y Derecho. Fragmentos para un derecho antisubordiscriminatorio,* Santiago de Chile, Ed. Olejnik, 2019, pp. 287-289.

alguna forma, obstaculiza el avance del Derecho antidiscriminatorio y que, como trataré de mostrar, no facilita la articulación de respuestas frente a la discriminación algorítmica.

De acuerdo con las Directivas europeas sobre la materia, por discriminación directa se entiende aquella situación en la que una persona es tratada de manera menos favorable que otra que se halla en situación análoga en función de un rasgo considerado prohibido[56]. Pensemos, por ejemplo, en el conocido caso de la compañía Amazon, que, después de utilizar durante más de 4 años un algoritmo en sus procesos de selección, procedió a eliminarlo al comprobar que contenía un sesgo implícito que penalizaba a las mujeres candidatas a un puesto de trabajo[57].

El concepto de discriminación directa es aparentemente sencillo, dado que hace referencia a una característica definitoria de una persona que motiva un trato discriminatorio. Sin embargo, para las teorías críticas –como la óptica iusfeminista– es una categoría difícil, ya que, por una parte, exige una revisión permanente del principio de igualdad de trato cuya vulneración da lugar a la discriminación y, por otra, depende de la idea de justicia de quien emite el juicio sobre la igualdad y la desigualdad y selecciona los patrones de comparación.

Cuando se hace referencia a la discriminación directa, se subrayan las tres dimensiones que la conforman: la intencionalidad, la simetría y el térmi-

[56] "Se considerará que existe discriminación indirecta cuando una disposición, criterio o práctica aparentemente neutros sitúen a las personas que tengan una determinada religión o creencia, una determinada discapacidad, una determinada edad o una determinada orientación sexual en una situación de desventaja particular con respecto a otras personas, a menos que dicha disposición, criterio o práctica se justifique objetivamente con una finalidad legítima y que los medios para alcanzar dicha finalidad sean adecuados y necesarios". Directiva sobre igualdad de trato entre hombres y mujeres en el acceso a los bienes y servicios, artículo 2(a). Directiva 2006/54/CE del Parlamento Europeo y del Consejo, relativa a la aplicación del principio de igualdad de oportunidades e igualdad de trato entre hombres y mujeres en asuntos de empleo y ocupación (refundición) (5 de julio de 2006). Directiva 2004/113/CE por la que se aplica el principio de igualdad de trato entre hombres y mujeres al acceso a bienes y ser vicios y su suministro (13 de diciembre de 2004). Directiva 2000/78/CE del Consejo, relativa al establecimiento de un marco general para la igualdad de trato en el empleo y la ocupación (27 de noviembre de 2000) Directiva 2000/43/CE del Consejo, relativa a la aplicación del principio de igual-dad de trato de las personas independientemente de su origen racial o étnico (29 de junio de 2000).

[57] J. DASTIN, "Amazon scraps secret AI recruiting tool that showed bias against women", *Reuters*, 8 de octubre de 2019. Disponible en: *https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.* Acceso: 08/10/2019.

no de comparación. La intencionalidad y la simetría han sido siempre conceptos especialmente problemáticos para definir la discriminación directa e indirecta[58]. En relación con la intencionalidad, se ha cuestionado su deslizamiento o su transformación en un elemento de justificación, su peso en los diversos niveles de examen[59], así como en los debates sobre la prueba. Una de las implicaciones de la vinculación de la intencionalidad con la discriminación directa ha sido la asunción de que la discriminación algorítmica es, prácticamente en todos los casos, una propiedad emergente no intencional del uso de los algoritmos y que, por tanto, ha de ser tratada como una discriminación indirecta. Pero, como sostienen Barocas y Sbelt, esta inferencia no facilita demasiado las cosas[60]. Constituye, más bien, una dificultad.

Un segundo elemento controvertido de la discriminación directa es el término de comparación. Considerado un elemento básico del juicio de igualdad y no discriminación en la cultura jurídica, se trata, en realidad, de un concepto resbaladizo que, como acabo de señalar, impone una revisión constante del principio de igualdad de trato desde un enfoque crítico como la óptica iusfeminista. La construcción del estándar de comparación adecuado, real o hipotético, respecto al cual se establece la desventaja sufrida por la persona demandante es uno de los aspectos más debatidos a la hora de analizar la capacidad del Derecho antidiscriminatorio para dar cabida a formas de desigualdad compleja –en terminología de Morondo–, entre ellas la interseccionalidad, la discriminación algorítmica[61] y las concepciones más estructurales de la desigualdad del Derecho antidiscriminatorio[62].

---

58 D. MORONDO, "Desigualdad compleja e interseccionalidad: "reventando las costuras" del Derecho antidiscriminatorio", en *Desigualdades complejas e Interseccionalidad. Una revisión crítica*, Madrid, IISJ-Dykinson, 2020, pp. 18-1.; y S. FREDMAN, "Direct and Indirect Discrimination. Is There Still a Divide?", en H Collins y T. Khaitan (eds.), *Foundations of Indirect Discrimination Law,* Portland, Hart Publ., pp. 31-55.

59 D. MORONDO, "Desigualdad compleja e interseccionalidad…", cit., p. 19. L. RONCONI, "Repensando el principio de igualdad: alcances de la igualdad real", *Isonomia,* num. 49, 2019, pp. 103-140.; y A. RIEKE, M. BOGEN y D ROBINSON, "Public scrutiny of automated decisions: early lessons and emerging methods" Upturn and O midyar Network, 2018, p. 25.

60 S. BAROCAS y A.D. SELBST "Big Data's Disparate Impact", cit., pp. 702, 711 y 723-730.

61 D. MORONDO "Desigualdad compleja e interseccionalidad", cit., p. 20.; y S. ATREY, "Comparison in Intersectional Discrimination", *Legal Studies,* num. 38 vol. 3, 2018, 379-395. *https://doi.org/10.1017/lst.2017.17*

62 De ahí que se presenten propuestas para conceptualizar y probar la discriminación a través de otro tipo de razonamiento que no sea el comparativo. En este sentido, es recomendable la lectura de S. GOLDBERG, "Discrimination by comparison", *Yale Law Journal*,

El término de comparación es un elemento que enlaza la dimensión conceptual de la discriminación con la cuestión fundamental de su prueba. La discriminación directa podría ser probada en aquellos casos en los que la pertenencia al grupo explícitamente incluido en la norma determine de manera automática un resultado negativo. Ahora bien, como afirma Soriano[63], no suele ser fácil acreditar este extremo en el ámbito probatorio, dado que es posible que no haya una referencia expresa al grupo desaventajado y que podrían utilizarse otros patrones o datos que establecen correlaciones o que no identifican el motivo discriminatorio. En muchos casos no se puede acceder al modelo algorítmico o conocer las variables empleadas y, por tanto, no es posible saber si se ha tomado en consideración un motivo prohibido o no. Así, los estereotipos y los sesgos introducen barreras relevantes en la prueba de la discriminación y en la determinación de un término de comparación[64].

Por su parte, la discriminación indirecta consiste en aquella situación en la que una disposición, práctica o criterio aparentemente neutros sitúan a una persona o grupo en una situación de desventaja respecto de otra sobre la base de un rasgo o categoría prohibida[65]. Si, como se afirma, el algoritmo

---

num. 120, 2011, pp. 779-790; R. HOLTMAAT, "De igual tratamiento a igual derecho", en D. HEIM y E. BODELÓN (eds.), *Derecho, Género e Igualdad. Cambios en las estructuras jurídicas androcéntricas*, Barcelona, Universitat Autònoma de Barcelona, 2010, pp. 209-228; y M.J. AÑÓN "Transformations in anti-discrimination law: progress against subordination", *Revus. Journal for Constitutional Theory and Philosophy of Law*, num. 40, 2020, pp. 27-43.

63 Soriano lleva a cabo un análisis exhaustivo de la prueba tanto en los supuestos de discriminación directa como en los de discriminación indirecta en el que profundiza en sus posibilidades y sus límites (A. SORIANO, "Decisiones automatizadas y discriminación…", cit., pp. 17-23, 30). En el texto –cuya lectura es muy recomendable–, la autora explica que no solo es un obstáculo la inteligibilidad y opacidad de los sistemas y que, por tanto, es probable que ello impida considerar discriminatorio de plano un sistema sin atender a elementos de comparación que tendrán que proceder de los datos anonimizados que se hayan utilizado o porque se puedan realizar experimentos con el programa, introduciendo diferentes perfiles para comprobar los resultados con unos y otros; sugiere, además, que si nos guiamos por la jurisprudencia del TJUE –por ejemplo, en materia de igualdad retributiva entre mujeres y hombres– y la extrapolamos al ámbito que aquí analizamos, la parte demandada deberá probar que la política salarial no es discriminatoria. Pero si el demandante no puede aportar un término de comparación, la parte demandada no tendrá la obligación de explicar la lógica subyacente a la decisión –es decir, no tendrá que revelar su código fuente.

64 D. MORONDO y E. GUIDONI, "El papel de los estereotipos en las formas de la desigualdad compleja: algunos apuntes desde la teoría feminista del derecho antidiscriminatorio", *Discusiones,* num. 28 vol. 1, 2022.

65 Aunque se trata de un concepto ampliamente conocido, reproduzco aquí el tenor literal del artículo 2 b) de la Directiva europea 2006/54/CE: "La situación en la que una dispo-

constituye la disposición aparentemente neutra, se trata de determinar las variables específicas utilizadas por los sistemas algorítmicos que dan lugar a un resultado discriminatorio.

Como señala Soriano, una de las implicaciones más espinosas de dar por descontado que la mayoría de los casos de discriminación algorítmica serán tratados como supuestos discriminación indirecta es que se trata de una modalidad de discriminación que permite un mayor margen de justificación[66]. En la discriminación indirecta, el estándar justificativo consta de dos pasos: el propósito o la finalidad legítima de la norma y la superación del juicio de proporcionalidad a través de los subprincipios de idoneidad, necesidad y proporcionalidad en sentido estricto. En muchas ocasiones, este examen se sintetiza considerando legítima toda finalidad lícita que no exprese en sí misma un ánimo de discriminar o que no trate de ocultarlo de manera evidente, pero en el ámbito de la discriminación por impacto o efecto esta interpretación resulta difícilmente aplicable[67].

Estas consideraciones llevan a preguntarse sobre pretensión de objetividad de la justificación de la discriminación, es decir, a plantear la cuestión de si hay una presunción de objetividad y neutralidad en el uso de modelos predictivos algorítmicos. Para dar respuesta a esta cuestión, parece imprescindible situar el análisis de las modalidades de discriminación "directa" e "indirecta" en un marco interpretativo más amplio que ha de retomar el sentido o los sentidos y la finalidad del Derecho antidiscriminatorio, extremos sobre los que ha habido desacuerdos doctrinales desde sus orígenes; estas divergencias se han centrado especialmente en la capacidad del Derecho antidiscriminatorio para lograr la transformación social o el cambio estructural[68].

Ahora bien, este ámbito se enfrenta una tensión irresuelta entre dos concepciones del Derecho antidiscriminatorio. Por una parte, el enfoque basa-

---

sición, criterio o práctica aparentemente neutros sitúan a personas de un sexo determinado en desventaja particular con respecto a personas del otro sexo, salvo que dicha disposición, criterio o práctica pueda justificarse objetivamente con una finalidad legitima y que los medios para alcanzar dicha finalidad sean adecuados y necesarios".

66 A. SORIANO, "Decisiones automatizadas y discriminación…", cit., pp. 23 y ss.

67 Ibid., pp. 20 y ss. En este caso, señala la autora, aceptada una finalidad legítima cabe preguntarse si el algoritmo es adecuado/idóneo para predecir el fenómeno del que se ocupa; por tanto, lo relevante es la mayor o menor exactitud predictiva del sistema. La parte demandada habrá de probar que no hay un modelo algorítmico menos discriminatorio o una base de datos más completa o menos sesgada que eventualmente podrían haberse utilizado.

68 D. MORONDO, "Desigualdad compleja…", cit., p. 17.

do en la igualdad de trato entendida como trato no arbitrario o razonable, que otros autores denominan principio de igualdad "anticlasificación"[69], y que, en cierto modo, presupone el trato indiferenciado no discriminatorio. Por otra, la perspectiva que atribuye al Derecho antidiscriminatorio la finalidad de hacer frente a la exclusión social, a la opresión y a la "subordiscriminación"[70].

Como señala Saba[71], la primera perspectiva parte de la intuición de que la igualdad exige un trato no arbitrario o, según la jurisprudencia de tribunales de referencia, un trato razonable. Esta concepción de la igualdad rechaza cualquier trato diferenciado entre las personas que pueda considerarse arbitrario, en el sentido de que el criterio que justifique ese trato no guarde una relación funcional con el fin perseguido. Este enfoque no considera relevante la existencia personas y grupos que han sido histórica y sistemáticamente excluidos en diversos ámbitos de la vida social, puesto que presupone por diversas razones que se dan ciertas condiciones de igualdad de oportunidades y de no sometimiento de algunos grupos (en el sentido de trato desigual grupal histórico, sistemático y, por ello, estructural)[72]. Esta concepción, actualmente dominante en el Derecho antidiscriminatorio, entiende la discriminación como la quiebra de la igualdad de trato en término básicamente individuales, y asume una la visión litigante y una concepción reparadora de la discriminación centrada en la dicotomía discriminación directa/indirecta y sus elementos nucleares –es decir, el término de comparación, la intencionalidad, la justificación objetiva de la discriminación y la litigación–.

La segunda concepción del Derecho antidiscriminatorio se articula en torno a la noción de la igualdad entendida como no sometimiento o no exclusión[73]. En ella se integran corrientes que tratan de que el Derecho anti-

[69] Barocas y Selbst denominan ambas concepciones "Derecho anticlasificación" y "Derecho antidiscriminación" (S. BAROCAS y A.D. SELBST, "Big Data's Disparate Impact", cit., p. 723).

[70] M. A. BARRÈRE, "Filosofías del Derecho Positivo ¿Qué Derecho y qué discriminación? Una visión contra-hegemónica del Derecho Antidiscriminatorio", *Anuario de Filosofía del Derecho,* vol. XXXIV, 2018, pp. 28-29.

[71] R. SABA, "Desigualdad estructural y acciones afirmativas", en A. VARAS y P. DÍAZ-ROMERO (eds.), *Acción afirmativa. Política para una democracia efectiva*, Santiago de Chile, Fundación Equitas-Ril Editores, 2013, pp. 85-92.

[72] Sobre esta distinción, véase H. NORTON, "The Supreme Court's Post-Racial Turn Towards a Zero-Sum Understanding of Equality", 52 *William & Mary Law Review*, 197, 2010, 206-215.

[73] R. SABA, "Desigualdad estructural y acciones afirmativas", cit., pp. 93-94.

discriminatorio no se limite a dar respuesta a demandas individuales (incluso grupales) de discriminación y configure un entramado normativo que asuma como objetivos la transformación de las estructuras de desigualdad y la emancipación de los sujetos –por ejemplo, a través de la introducción de nuevas categorías como la discriminación por asociación, múltiple, interseccional o la perspectiva interpretativa sistémica–. Se trata de un enfoque orientado a identificar la plasmación en los sistemas jurídicos de aquellas desigualdades que tienen su razón de ser en las estructuras de poder con capacidad para ordenar las relaciones sociales, para atribuir o asignar estatus (subordinados o inferiores, privilegiados o superiores) y para establecer dinámicas e inercias que reproducen estas relaciones de subordinación[74]. Partiendo de esta premisa, Mackinnon afirma que el principio de igualdad sustantiva no es una abstracción, sino el contenido social específico de cada "desventaja preexistente" basada en motivos concretos, unas desventajas cuyo principio identificador es la jerarquía social. En otros términos, la jerarquía sobre bases concretas es lo que provoca la desigualdad[75]. Precisamente por ello, la autora considera que la discriminación indirecta (por impacto o por efecto) fracasa repetidamente o muestra sus insuficiencias porque "a menos que se sepa que una disparidad es el impacto o el efecto de una jerarquía sustantiva preexistente, se puede eludir su tratamiento, y eso se hace repetidamente". Como ha mostrado de forma reiterada Barrère, entre muchas otras voces, ningún sistema jurídico existente "se libra de los efectos de la exclusión originaria de las mujeres", es decir, del falso universalismo; ninguno de ellos la ha afrontado directamente y, en términos generales, las respuestas jurídicas han sido insuficientes para responder ante este tipo de exclusión[76]. Este déficit interpela a la visión dominante de la discriminación indirecta, que legitima la tesis de acuerdo con la cual en ella solo hay una situación puntual o accidental que puede afectar indistintamente a las personas y los grupos subordinados y a quienes dominan y oprimen, y que sostiene que puede justificarse "objetivamente"[77]. En la arena jurídica, la objetividad se

[74] M.A. BARRÈRE, "Filosofías del Derecho Positivo…", cit., p. 32.

[75] C. MacKINNON, "Substantive equality revisited: A reply to Sandra Fredman", International Journal of Constitutional Law, num. 14 vol. 3, 2016, pp. 742 y 744.

[76] Barrère ha abordado la dimensión del falso universalismo en diversos trabajos sobre feminismo jurídico. Por todos, remito a M.A. BARRÈRE, *El Derecho Antidiscriminatorio y sus límites. Especial referencia a la perspectiva iusfeminista,* Lima, Grijley, 2014, pp. 13-17.

[77] M.A. BARRÈRE, "La igualdad de género desde el activismo de las profesiones jurídicas", en *Feminismo y Derecho. Fragmentos para un derecho antisubordinación*, cit., p. 261.

identifica muy a menudo con la neutralidad, aunque sabemos que ni la legitimidad o finalidad de la norma ni la necesidad o idoneidad de los medios responden a objetivos neutrales. Parece, por tanto, que la fertilidad de la discriminación indirecta se encuentra en la segunda perspectiva del Derecho antidiscriminatorio.

Sin embargo, es preciso subrayar que aún no se ha producido un compromiso claro del Derecho antidiscriminatorio con el principio de antisubordinación[78] y que es claramente dominante la perspectiva que concibe el principio de igualdad de trato como indiferenciación o trato razonable y no arbitrario (principio de anticlasificación) y lo identifica con la prohibición de la discriminación intencional (explicitada al seleccionar características de las clases protegidas que producen un daño) o implícita/encubierta. En el ámbito del Derecho antidiscriminatorio, también se habla de evaluar la conveniencia de ampliar las características protegidas, incorporando rasgos como la pobreza y las desventajas socioeconómicas o lo efectos que provoca en las personas y los grupos el desarrollo tecnológico[79]. A pesar de que la discriminación indirecta adolece de cierta insuficiencia constitutiva que funciona como una barrera para hacer frente a la discriminación estructural, se propone como paraguas jurídico para cobijar la discriminación digital.

Por ello, y ante las dificultades que plantea la discriminación algorítmica, se ha propuesto la reconducción de la discriminación hacia modalidades distintas a la discriminación indirecta –específicamente, la discriminación por asociación y la discriminación interseccional–, dado que se considera que estas categorías pueden superar algunas de las insuficiencias de aquella. Si se amplían los argumentos que permitan analizar mejor los datos, el sistema y, sobre todo, las correlaciones tal vez será posible encontrar aquí otros asideros en la medida en que se amplíe el rango de argumentos que permitan analizar mejor los datos, el sistema y, sobre todo, las correlaciones. Sin embargo, el núcleo de la cuestión sigue estando donde lo hemos situado: en la finalidad del Derecho antidiscriminatorio y en las cuestiones de prueba. La discriminación algorítmica precisa ser abordada desde la perspectiva del Derecho antidiscriminatorio nucleada en torno a la antisubordinación o la antiexclusión. En este sentido, es probable que la noción discriminación algorítmica interseccional tenga un mayor potencial para generar avances en

[78] S. BAROCAS y A.D. SELBST, "Big Data's Disparate impact…", cit., p. 724.

[79] J. GERARDS y F. ZUIDERVEEN, "Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making…", cit., p. 62 y ss.

materia de prueba si esta puede practicarse teniendo en cuenta las relaciones de correlación.

Por otro lado, respecto a la tesis de que la discriminación indirecta puede justificarse mediante razones objetivas, se apela a su apoyo en la "objetividad" de los datos o en la posibilidad de analizar el sistema de inteligencia artificial. Pero esta objetividad no puede darse por supuesta fundamentalmente porque lo que se define como datos es "el estado actual del mundo".

En este sentido, Solar Cayón da cuenta de algunas propuestas orientadas a fundamentar la pretensión de objetividad[80], pretensión centrada en el examen de los resultados de aprendizaje de los algoritmos a partir de datos e indicadores matemáticos. El autor hace referencia a tres vías: (a) las estrategias anticlasificación, que exigen la no inclusión en los datos de rasgos o atributos prohibidos (por ejemplo, raza o género) y también sus *proxies* (es decir, atributos correlacionados con aquellos, por ejemplo el código postal), (b) las estrategias de paridad clasificatoria, que establecen criterios para verificar que el rendimiento predictivo del sistema es igual para los diversos grupos definidos por los rasgos jurídicamente protegidos; esta estrategia impediría tomar como datos objetivos "el estado actual del mundo", y (c) las estrategias de calibración, orientadas a comprobar si los resultados son independientes de los atributos protegidos.

Sin embargo, las orientaciones basadas en el conocimiento estadístico y matemático son insuficientes. Como apunta el propio Solar Cayón[81], una parte significativa de la doctrina aboga por la adopción de un enfoque amplio que, junto a los aspectos técnicos y estadísticos, tenga en cuenta el contexto social, institucional y jurídico, así como el punto de vista de los distintos grupos que pueden verse afectados. Esta perspectiva explica el cuestionamiento –que, entre otros, propone Soriano[82]– de las "condiciones de fondo" en las que se toman las decisiones sobre la clasificación y selección de individuos, dado que, como explica la autora, son estas condiciones las que permiten entender que, incluso allá donde la discriminación es suficientemente precisa y aparentemente eficiente, es el resultado de las relaciones e interacciones que tienen lugar en una sociedad que sistemáticamente sitúa a determinados grupos en posición de desventaja o desigualdad. Por ello, si no se superan o transforman estas situaciones de discriminación aparentemente eficientes,

[80] J.I. SOLAR CAYÓN, "Inteligencia artificial en la justicia penal", cit., pp.160-161.
[81] Ibid.,165.
[82] A. SORIANO, "Decisiones automatizadas y discriminación…", cit., pp. 22-29.

no será posible modificar las estructuras que sitúan en un punto de partida desigual a las personas pertenecientes a grupos desaventajados[83].

De lo dicho hasta aquí cabe concluir que el Derecho antidiscriminatorio –especialmente, la segunda concepción que he expuesto (antisubordinación)– cobra especial importancia ante el desarrollo algorítmico, en la medida en que está llamado a ofrecer respuestas para abordar graves desigualdades estructurales. Este enfoque conduce, así, a una perspectiva más amplia que converge con una noción sustantiva de la justicia social vinculada con los derechos humanos capaz de dar cuenta de los fundamentos de las normas concretas que regulen la aplicación de la inteligencia artificial, de identificar los riesgos que las decisiones algorítmicas pueden provocar en los derechos fundamentales y de cuestionar si aquellas normas amenazan principios jurídicos relevantes[84].

4. SOBRE EL VALOR DE LA EFECTIVIDAD PREDICTIVA

Pensemos en aquellos supuestos en los que el uso de inteligencia artificial genera unos resultados discriminatorios que, sin embargo, son eficientes de acuerdo con el modelo que se propone, tienen exactitud predictiva. En este sentido, no bastaría argumentar que el efecto discriminatorio es resultado de un prejuicio irracional –y, por tanto, injusto–, del mismo modo que sería insuficiente limitarse a señalar el carácter erróneo o sesgado de los datos seleccionados o localizar un ejemplo de mal etiquetado o un conjunto insuficiente de características. En este caso, los argumentos que justifican las prohibiciones tendrían que incorporar un compromiso sustantivo antidiscriminatorio sobre las desigualdades estructurales y, por tanto, deberían estar basados en el principio de antisubordinación. La disyuntiva es clara:

[83] Janneke y Zuiderveen subrayan esta dimensión o proyección de la discriminación algorítmica en el refuerzo de las desigualdades estructurales (G. JANNEKE, F. ZUIDERVEEN, "Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making…", cit.).

[84] F. ZUIDERVEEN BORGESIUS, "Strengthening legal protection against discrimination by algorithms and artificial intelligence", *The International Journal of Human Rights,* num. 4 vol. 10, 2020, pp. 1584-1586. DOI: 10.1080/13642987.2020.1743976. El autor considera que el derecho antidiscriminatorio y la legislación de protección de datos en el ámbito europeo son instrumentos jurídicos adecuados para luchar contra la discriminación algorítmica y sugiere algunas mejoras en su aplicación que pasan por el desarrollo de legislaciones sectoriales que regulen los riesgos que comporta para cada uno de los derechos.

"cuanto más se aleja la doctrina de la reparación de la reparación sustantiva, menos utilidad tiene para remediar este tipo de efectos discriminatorios"[85]. Considero que una teoría de corte más sustantivo ha de contar al menos con algunas piezas básicas: por una parte, una concepción de la igualdad y la no discriminación que aborde en profundidad la existencia de estructuras sociales de opresión o subordinación; por otra, una vinculación expresa entre el reconocimiento y garantía del derecho a la igualdad y no discriminación y la protección de la libertad entendida como la capacidad real de participar en la toma de decisiones y como libertad material.

Se trata, en definitiva –y este es un camino que aún está por recorrer–, de suministrar argumentos orientados a establecer límites a determinados desarrollos de la inteligencia artificial que erosionan los derechos humanos, que provocan un incremento de las desigualdades o que exacerban las discriminaciones existentes. En otros términos, el objetivo es en suma, de garantizar los principios de igual-libertad, equidad, rendición de cuentas, inteligibilidad y transparencia. En este sentido, Grace y Bamford proponen emplazar la gobernanza algorítmica, sus presupuestos, modelos, sistemas o procesos, así como su supervisión –transparencia y responsabilidades– bajo el parámetro de los derechos humanos y la justicia[86]. Las autoras sostienen que una teoría igualitarista de la justicia como la de Rawls resulta fecunda para hacer frente a las desigualdades digitales, dado que podría orientar la gobernanza digital mediante la regulación de las instituciones sociales y económicas básicas de acuerdo con los principios de justicia. Consideran, pues, que un marco teórico tan amplio como la propuesta rawlsiana es capaz de brindar una cobertura justificatoria a la denominada gobernanza algorítmica.

Grace y Bamford parten de los principios de justicia de Rawls. El primer principio ("Cada persona ha de tener un derecho igual al esquema más extenso de libertades básicas que sea compatible con un esquema semejante de libertades de los demás"[87]) ofrece las bases para derivar legítimamente el imperativo de crear vías accesibles para examinar y controlar la incidencia sobre esa igual libertad –por ejemplo, saber si se ha elaborado un perfil de una persona, entender los pasos que ha dado un modelo algorítmico para llegar a determinada conclusión o pedir cuentas a quienes gobiernan de las

[85] S. BAROCAS y A.D. SELBST, "Big Data's Disparate Impact…", cit., p. 728.

[86] J. GRACE y R. BAMFORD, "AI Theory of Justice: Using Rawlsian Approaches …", cit., p. 2.

[87] J. RAWLS, *Teoría de la justicia,* Madrid, Fondo de Cultura económica, 2002, p. 67.

bases de conocimiento de sus decisiones–. Por tanto, el primer principio se proyecta fundamentalmente en la transparencia, la accesibilidad al conocimiento, el acceso a la justicia, las salvaguardas respecto a las decisiones totalmente automatizadas. Este enfoque exige la disponibilidad de un alto grado de información que permita a los individuos impugnar las decisiones por diversos motivos, dado que la falta de transparencia de los algoritmos que se utilizan para adoptar decisiones gubernamentales constituye una amenaza innata para un sistema de libertades iguales para todos. Esto afecta tanto a los criterios que se tiene en cuenta como al riesgo que puede comportar la pretensión de exactitud predictiva. Los criterios y elementos que se seleccionan han de respetar tanto el principio de publicidad como el de ser sensibles a las situaciones y condiciones en las que se sitúan los sujetos. Ambos son necesarios para que no se reproduzca la injusticia estructural. En cuanto el riesgo, no tiene más contrapeso que la transparencia.

El segundo principio ("Las desigualdades sociales y económicas habrán de ser estructuradas de manera que sean para (a) beneficio de los menos aventajados de acuerdo con un principio de ahorro justo y (b) unidos a los cargos y las funciones asequibles a todos, en condiciones de justa igualdad de oportunidades"[88]), el denominado principio de la diferencia, se presenta como el fundamento para garantizar la igualdad real y efectiva en el ámbito de la inteligencia artificial a través de algunas vías, entre las que destacan la legislación primaria de desarrollo del principio de igualdad (sustantiva) y la prohibición de discriminación, así como la imposición a los poderes públicos de obligaciones ligadas al respeto de los derechos humanos[89]. No obstante, las citadas autoras reconocen que probablemente sería necesaria una regulación adicional para proteger la justicia y los derechos humanos que, sin embargo, no concretan. En este punto, la propuesta me parece insuficiente para dar respuesta a la cuestión clave de la que venimos hablando: la desigualdad estructural.

Sin embargo, considero que el segundo principio adquiere más vigor si lo examinamos a la luz de otra exigencia: la eficiencia, esto es, la exactitud predictiva. De acuerdo con la concepción rawlsiana, si la estructura básica de la sociedad no respeta los principios de iguales libertades básicas e igualdad de oportunidades, la justicia exige corregir dicha situación, aunque ello

[88] Ibid., p. 280.

[89] M.J. AÑÓN, "Human rights obligations, specially, in times of crisis", *The Age of Human Rights Journal*, vol. 17, 2021, pp. 1-26.

comporte cambios que afecten a la eficiencia. Como explica Bayón[90], en esta concepción la eficiencia es relativa a alguna distribución inicial, teniendo en cuenta que el principio de justicia distributiva es el principio de la diferencia y que la eficiencia se situaría en el punto de partida de una estructura básica que respete los anteriores principios. En este sentido, la teoría de la justicia integra o es compatible con la eficiencia si nos encontramos en un contexto que respeta los principios de iguales libertades básicas y la igualdad de oportunidades y en el que solo son admisibles las desigualdades materiales que maximizan las perspectivas de los menos favorecidos. Si se satisface el principio de la diferencia, no es posible mejorar la situación de alguien sin empeorar la de algún otro. La eficiencia, en ese supuesto, es un ingrediente de la justicia supeditada a la satisfacción previa de ciertas condiciones básicas que no sería legítimo sacrificar para su salvaguarda.

5. NOTAS CONCLUSIVAS

La preocupación central de este artículo ha girado sobre la preservación de los derechos humanos, sus garantías y el principio de eficiencia –la pretensión de exactitud predictiva– en el uso de modelos algorítmicos de decisión.

Hemos podido constatar que la inclusión del uso de inteligencia artificial en el marco interpretativo del riesgo o de la prevención de daños a bienes y derechos fundamentales, que parece ser la perspectiva que se propone en el Derecho de la Unión Europea, no zanja la cuestión de la racionalidad decisoria. De otro lado, vincula las obligaciones orientadas a esta pretensión a una obligación genérica de diligencia debida en el marco de los límites normativos ya existentes relativos a las garantías y protección de los derechos fundamentales.

Entre las conductas consideradas de alto riesgo tienen una posición privilegiada aquellas que pueden dar lugar a resultados discriminatorios básicamente porque existen sesgos en los datos y/o en la clasificación. La forma hegemónica de tratar esta cuestión ha reconducido la discriminación algorítmica a las categorías bien conocidas de discriminación directa/indirecta integradas en el Derecho europeo y que han cristalizado en un esquema binario que, de alguna forma, no facilita la articulación de respuestas frente a la discriminación algorítmica.

[90] J.C. BAYÓN, "Justicia y Eficiencia", en ALMOGUERA *et al.*, *Estado, justicia, derechos,* Madrid, Alianza, 2002, pp. 258-259.

El articulo da cuenta, en este sentido, de una tensión irresuelta entre dos concepciones del Derecho antidiscriminatorio y opta por una perspectiva que atribuye al Derecho antidiscriminatorio la finalidad de hacer frente a la exclusión social, a la opresión y a la "subordiscriminación".

Esta concepción del Derecho antidiscriminatorio cobra especial importancia ante el desarrollo algorítmico, en la medida en que está llamado a ofrecer respuestas para abordar graves desigualdades estructurales. Este enfoque propicia la adopción de una perspectiva más amplia que converge con una noción sustantiva de la justicia social vinculada con los derechos humanos capaz de dar cuenta de los fundamentos de las normas concretas que regulen la aplicación de la inteligencia artificial, de identificar los riesgos que las decisiones algorítmicas pueden provocar en los derechos fundamentales y de cuestionar si aquellas normas amenazan principios jurídicos relevantes.

María Jose Añón Roig
Facultad de Derecho
Universidad de Valencia
Avda. dels Tarongers, s/n. Edificio Occidental
46071 Valencia
e-mail:Maria.J.Anon@uv.es