INTELIGENCIA ARTIFICIAL Y SESGOS DISCRIMINATORIOS: ¿ES NECESARIO UN NUEVO CONCEPTO DE DISCRIMINACIÓN ALGORÍTMICA?

Artificial intelligence and discriminatory bias: Is a new concept of algorithmic discrimination necessary?

JOSÉ FERNANDO LOUSADA AROCHENA
Universidad de A Coruña
fernando.lousada@udc.es

Cómo citar/Citation
Lousada Arochena, J. F. (2024).
Inteligencia artificial y sesgos discriminatorios:
¿es necesario un nuevo concepto de discriminación algorítmica?
IgualdadES, 11, 97-123
doi: https://doi.org/10.18042/cepc/IgdES.11.04

(Recepción: 12/07/2024; aceptación tras revisión: 07/10/2024; publicación: 13/12/2024)

Resumen

Los sistemas de IA incurren en sesgos discriminatorios y, en particular, en sesgos de género, que presentan diferencias respecto a los sesgos existentes con carácter general en las personas y en la sociedad. Los sesgos en la IA tienden a incrementar los sesgos generales y, en todo caso, presentan manifestaciones específicas. La cuestión que se aborda en el presente estudio es si los conceptos jurídicos vigentes sobre discriminación son suficientes para dar protección frente a tales sesgos o no, y en ese caso si se haría necesaria la creación de un nuevo concepto de discriminación algorítmica. La respuesta a la cuestión exige analizar las causas de los sesgos discriminatorios en los sistemas de IA, sus manifestaciones y las dificultades de calificación jurídica. Después se analizará lo que aporta a la cuestión el reglamento europeo de inteligencia artificial. Finalizamos con unas consideraciones que no pretenden ser conclusiones cerradas, sino propuestas para el debate.

Palabras Clave

Discriminación; sesgos de género; inteligencia artificial.

Abstract

AI systems incur discriminatory biases and, in particular, gender biases, which present differences with respect to the biases that generally exist in people and in society. Biases in AI tend to increase general biases and, in any case, present specific manifestations. The question to be addressed in this study is whether the current legal concepts on discrimination are sufficient to provide protection against such biases, or not, and in that case whether the creation of a new concept of algorithmic discrimination would be necessary. The answer to the question requires analyzing the causes of discriminatory biases in AI systems, their manifestations and the difficulties of legal qualification. Afterwards, what the European Artificial Intelligence Act contributes at the issue. Finally, some considerations that are not intended to be closed, but rather proposals for debate.

Keywords

Discrimination; gender bias; artificial intelligence.

SUMARIO

I. INTRODUCCIÓN

IA y sesgos discriminatorios son dos tópicos frecuentemente relacionados. Queda ya lejana, cuando menos a nivel de conocimiento científico (aunque aún pervive con fuerza en el imaginario popular), la inocente idea de que, como los sistemas de IA se sustentan en un algoritmo, los resultados de su aplicación siempre tendrían tanto la exactitud como la objetividad característica de las matemáticas, sin incurrir en la oblicuidad o torcimiento en que el sesgo consiste (oblicuidad o torcimiento de una cosa es la primera definición del sustantivo sesgo según el *Diccionario* de la RAE). Al contrario, la idea actualmente imperante, tanto en el ámbito de instituciones públicas y privadas (internacionales, europeas y nacionales) como de la doctrina científica, es que los sistemas de IA, en especial si son de aprendizaje automático *(machine learning)*, causan sesgos de alcance discriminatorio, en particular de género. En este sentido, la IA se ha tildado de «arma de destrucción matemática» que aumenta la desigualdad y amenaza la democracia (O'Neil, 2017: 2018), de «artefacto de alto riesgo para la igualdad de género» (Rodríguez, 2024: 25), con riesgo de que su uso por el Estado sin las debidas regulaciones le podría hacer perder «su carácter de Estado de Derecho, su condición de Estado democrático y la garantía del Estado social» (Presno, 2022: 89).

La cuestión que aborda este estudio es si todos los sesgos discriminatorios se pueden subsumir en los conceptos jurídicos existentes o estos son insuficientes, de manera que se haría necesaria la creación de un concepto de discriminación algorítmica.

Para responder a la cuestión damos por supuesto el conocimiento por el lector o lectora de los conceptos de discriminación contenidos en nuestra normativa antidiscriminatoria vigente, tanto al nivel de la Unión europea (en particular, directivas 2006/54/CE[1], 2004/113/CE[2], 2000/43/CEE[3] y 2000/78/CEE[4]) como a un nivel interno (en particular, Ley Orgánica 3/2007, de 22 de marzo[5], y Ley 15/2022, de 12 de julio[6]).

Y con la base de ese conocimiento, analizaremos con cierta minuciosidad las distintas manifestaciones de sesgos discriminatorios hasta el momento detectados con la finalidad de ofrecer una clasificación atendiendo a sus causas que permita identificar las dificultades de calificación jurídica de cada una de las categorías de sesgos (epígrafe 1).

Sobre la base de esta indagación, y dada la reciente aprobación del Reglamento de Inteligencia Artificial de la Unión europea (en adelante, el RIA)[7], es obligado verificar si ha supuesto modificación de los conceptos de discriminación contenidos en nuestra normativa vigente o si ofrece algún criterio para su interpretación (epígrafe 2).

---

1 Directiva 2006/54/CE del Parlamento Europeo y del Consejo, de 5 de julio de 2006, 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).

2 Directiva del Consejo 2004/113/CE, de 13 de diciembre de 2004, por la que se aplica el principio de igualdad de trato entre hombres y mujeres al acceso a bienes y servicios y su suministro.

3 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.

4 Directiva 2000/78/CE del Consejo, de 27 de noviembre de 2000, relativa al establecimiento de un marco general para la igualdad de trato en el empleo y la ocupación.

5 Ley Orgánica 3/2007, de 22 de marzo, para la Igualdad Efectiva de Mujeres y Hombres.

6 Ley 15/2022, de 12 de julio, Integral para la Igualdad de Trato y la No Discriminación.

7 Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) 300/2008, (UE) 167/2013, (UE) 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial).

Después de todos estos análisis estaremos en condiciones de realizar algunas consideraciones sobre la necesidad o no del concepto de discriminación algorítmica (epígrafe 3), un debate ya iniciado en la doctrina científica española con algunos aportes trascendentes (Sáez, 2020; Pérez, 2023; Todolí, 2024; Ginès, 2024), pero que en modo alguno está cerrado; al contrario, estamos aún en sus inicios y a él queremos contribuir.

II. MANIFESTACIONES DE SESGOS DISCRIMINATORIOS EN LOS SISTEMAS DE IA CLASIFICADAS SEGÚN SUS FUENTES Y DIFICULTADES DE CALIFICACIÓN JURÍDICA DE CADA UNA DE LAS CATEGORÍAS

Un sistema de IA es un artificio creado, desplegado y utilizado por personas. En consecuencia, los sesgos discriminatorios en los sistemas de IA provienen de los sesgos discriminatorios que, con carácter general, tenemos como personas y como sociedad. Pero esta respuesta tan genérica, siendo totalmente cierta en esos términos genéricos, puede conducir al engaño de creer que los sesgos discriminatorios en los sistemas de IA son idénticos, cuantitativa y cualitativamente, a los existentes con carácter general. Al contrario, hay causas diferenciales determinantes de que en los sistemas de IA se incrementen los sesgos discriminatorios existentes con carácter general. Y esas causas diferenciales superan lo cuantitativo pues determinan la aparición de manifestaciones específicas de sesgos discriminatorios. O sea, es un cambio también cualitativo, lo que obliga a abordar la calificación jurídica de esas nuevas manifestaciones desde la perspectiva de los conceptos preexistentes en el ámbito de la tutela antidiscriminatoria.

Para analizar las causas de los sesgos discriminatorios en los sistemas de IA, sus manifestaciones y las dificultades de calificación jurídica, los clasificaremos didácticamente según la fuente de la cual provengan: los datos usados por el sistema de IA (sesgos derivados de la mala calidad de los datos, sesgos por inferencia a partir de una variable *proxy*, sesgos por asimetría en las minas de datos y sesgos de retroalimentación), por el diseño del propio sistema de IA (sesgos en el algoritmo o en la interfaz, en especial si esta es inmersiva: metaverso) o los sesgos humanos (derivados de conductas humanas en el entorno del sistema de IA). A ellos se suman los sesgos de invisibilización, que, sin crear resultados sesgados, invisibilizan los que hay.

Advertir, en todo caso, que es una clasificación didáctica con la única finalidad de catalogar las diversas manifestaciones de sesgos discriminatorios para detectar las dificultades de subsunción de cada categoría en los conceptos preexistentes en el ámbito de la tutela antidiscriminatoria. Y es una

advertencia pertinente pues en la mayoría de casos los sesgos provienen de varias fuentes diferentes que interactúan y se potencian.

1. SESGOS EN LOS DATOS Y EN CÓMO SE PROCESAN: CAUSAS Y MANIFESTACIONES

La IA, como la inteligencia humana, utiliza información de entrada para llegar a decisiones, y la misma puede estar sesgada. Pero a partir de esta coincidencia esencial se aprecian causas diferenciales que propician cuantitativamente los sesgos discriminatorios en los sistemas de IA y que les otorgan sus características cualitativamente diferenciales más sobresalientes (y en ese sentido, los sesgos en los datos y en cómo se procesan son los sesgos más paradigmáticos en los sistemas de IA).

La primera de esas causas diferenciales se refiere a la capacidad de procesamiento, que es cuasi infinita en ciertos sistemas de IA (en especial, el *deeplearning* y la IA generativa), lo que permite la utilización de grandes cantidades de datos, incluyendo datos de estructura compleja *(big data)*. Y en estas circunstancias la mina de datos usada puede fácilmente incluir datos de mala calidad ética (datos discriminatorios, desactualizados, obtenidos o tratados sin garantía de cumplimiento de las leyes de protección de datos) sin que los sistemas de IA (salvo acaso los llamados algoritmos éticos, aunque, a día de hoy, no son los mayormente implantados) tengan la capacidad (que sí que tendría una decisión humana) de prescindir de los datos de mala calidad ética o de corregir decisiones por motivos éticos. Para expresar esta situación se suele aludir al acrónimo GIGO: «Garbage in = Garbage out» (si entra basura, saldrá basura) (Giralt, 2024: 10).

Veamos algunos ejemplos. En el ámbito del empleo son bastante habituales, en particular si se trata de empleo en plataformas, sistemas de IA que valoran la opinión de las personas usuarias del servicio, y aquí hay un riesgo importante de introducción de datos que incorporan sesgos discriminatorios, cuando no discriminaciones puras y duras, que el sistema no depura. Se ha determinado en algunos estudios realizados en Estados Unidos que cuando personas usuarias de Uber comprobaban que el conductor era de determinada raza a través de la propia aplicación, cancelaban el servicio y el algoritmo de Uber no identificaba que ello era una discriminación, sino que aprendía que esa cancelación se debía a falta de capacidad del conductor, con lo cual ese conductor recibía menos ofertas y finalmente era expulsado (Aragüez, 2024: 143).

Otras veces el sesgo se deriva de la calificación previa hecha por personas (y que por lo tanto puede contener un sesgo) de los datos que se le facilitan al sistema de IA. En este sentido, en Estados Unidos se ha denunciado la

denegación hasta hace poco de pruebas de cáncer de mama para las mujeres negras bajo la suposición de que solo eran necesarias para las mujeres blancas, y «esta denegación es, a su vez, una denegación de ratificación que, asimismo, influye en los programas computacionales de cálculo de riesgo» (IM, 2023: 21). Algo semejante se podría plantear en España dada la situación de infradiagnóstico de la EPOC, que afecta más a las mujeres (Soriano, 2021).

Incluso se ha llegado a plantear el riesgo de sesgo de sistemas donde los datos se obtienen de las propias personas. Cammio Xpress Analytics valora las características no profesionales de las personas candidatas a un empleo para verificar la capacidad de liderazgo, el sentimiento de pertenencia a la empresa y las posibilidades de negociar las condiciones. Se basa en la autoevaluación de «los cinco grandes» (OCEAN: *openness*, *conscientiousness*, *extraversion*, *agreableness*, *neuroticism*). Pero «los informes sobre uno mismo […] no solo muestran una autopercepción, sino que también contienen una percepción social implícita», de manera que estamos ante «una metodología que psicologiza la discriminación estructural y la toma no solo como un punto de partida descriptivo, sino normativo para justificar una decisión laboral» (IM, 2023: 22-23).

La segunda de esas causas diferenciales que propician los sesgos discriminatorios en los sistemas de IA se deriva de la capacidad de inferencia estadística de los sistemas de IA con aprendizaje automático. Aunque los datos manejados sean de calidad sin tacha de estereotipos o prejuicios, dichos sistemas pueden producir sesgos discriminatorios a partir del procesamiento de datos disponibles que son en principio objetivos: el código postal de una persona o el tiempo dedicado a leer unas u otras noticias le puede permitir al algoritmo deducir la raza, la religión o la ideología política. Un sistema de selección o promoción de personal puede interpretar la baja tasa de empleo femenino como una escasa habilidad de las mujeres para integrarse en la empresa. En suma, el algoritmo utiliza una variable próxima, en principio ajena a una causa de discriminación, para inferir una respuesta discriminatoria (es el sesgo discriminatorio por variable *proxy*).

Tal capacidad de inferencia estadística es la razón de ser de los algoritmos predictivos diseñados, desplegados y usados para inferir riesgos a partir de datos objetivos. De ahí que en esta clase de algoritmos se aprecian con intensidad los problemas de sesgos discriminatorios y de vulneración del derecho a la privacidad.

Veamos algunos ejemplos. Predective Hire identifica a aquellas personas trabajadoras con una mayor probabilidad de reclamar mejoras o de afiliación sindical a partir de datos obtenidos en las plataformas virtuales de la empresa. «Más allá de que se puedan ocasionar invasiones en la privacidad, la empresa tenía la capacidad de tomar decisiones y, en su caso, despedir a los trabajadores

por la información extraída del uso de las herramientas tecnológicas en los puestos de trabajo» (Aragüez, 2024: 145-146).

Compas valora el grado de reincidencia del individuo a partir de la información obtenida de una encuesta de más de cien preguntas y los antecedentes penales. Fue cuestionado en el caso *Loomis*, un condenado a seis años de prisión a quien la pena se le ajustó dado el riesgo de reincidencia tras aportar el fiscal un informe elaborado por el algoritmo Compas. El Tribunal Supremo de Wisconsin[8] no consideró vulnerado el derecho al debido proceso ni el principio de igualdad, a pesar de que Compas había sido puesto en duda alegando que reproducía sesgos raciales (*ibid.*: 154 ss).

En la Unión europea, el primer caso judicializado con respecto a los algoritmos predictivos de riesgo es el de SyRi. Fue desarrollado por el Gobierno holandés para predecir el fraude a la Seguridad Social a través del análisis de grandes cantidades de datos disponibles por las Administraciones públicas. La Corte de Distrito de La Haya, en su sentencia de 05/02/2020[9], consideró en general lícito utilizar instrumentos de este tipo si hay un interés público que lo justifique y se garantice la mínima injerencia en la privacidad; consideró que SyRI no ofrecía garantías suficientes. No entró a considerar si producía los sesgos discriminatorios que denunció Philip Alston, relator sobre la extrema pobreza y derechos humanos de Naciones Unidas, quien actuó como *amicus curiae* (Lazcoz y Castillo, 2020; Rivas, 2020: 354-356; Aragüez, 2024: 152 y ss.).

También en España se han analizado los sesgos causados por tres algoritmos predictivos: VioGen, para predecir el riesgo de agresión en casos de violencia de género; Riscanvi, para predecir el comportamiento en el ámbito carcelario, y VeriPol, para predecir los fraudes en denuncias de robo (Martínez *et al.*, 2024; Presno, 2023b). La problemática principal de estos algoritmos predictivos se encuentra en la ausencia de transparencia para comprobar las garantías de derechos de las personas y asimismo se ha cuestionado su efectividad en relación con el fin buscado de predicción de riesgos.

La tercera de esas causas diferenciales que propician los sesgos discriminatorios en los sistemas de IA, ligada con las dos anteriores, obedece a que en las minas de datos de las que se nutre el sistema pueden existir (y ello es habitual) marcadas asimetrías con incidencia sobre factores discriminatorios. Aunque se haya programado de manera neutra, el sistema aprende que las características

8 *Loomis v. Wisconsin*, 881 N.W.2d 749 (Wis. 2016), cert. denied, 137 S. Ct. 2290 (2017).

9 Sentencia del Tribunal del Distrito de La Haya, C/09/550982 / HA ZA 18-388, de 5 de febrero de 2020.

de las personas o grupos de los que hay sobreabundancia de datos son la normalidad a potenciar, mientras que las características de aquellas personas o grupos de los que hay carencia de datos son la anormalidad a excluir (es la «estadística contra las minorías» de la que nos habla Todolí, 2024: 64). Según un reciente estudio publicado por la OIT, la mayoría de los datos utilizados por los sistemas de IA proceden de las poblaciones occidentales, educadas, industrializadas, ricas y democráticas (WEIRD: *white*, *educated*, *industralized*, *rich and democratic* —literalmente, *weird* es raro—) (Gmyrek *et al.*, 2024). De ahí que la búsqueda en la red de trajes de novia muestra el clásico vestido blanco occidental, mientras que los trajes de novia de otras culturas se catalogan de vestidos típicos o folclóricos (Flores, 2023: 102).

Si observamos el fenómeno con perspectiva de género, también hay «una notable asimetría de datos», pues «los datos disponibles proceden o versan sobre los varones, siendo muchos menos los datos disponibles que proceden o versan sobre las mujeres», e incluso aún menos «en el caso de las mujeres racializadas, con discapacidad o de bajo nivel económico» (Rodríguez, 2024: 16). Los termostatos inteligentes en oficinas se convirtieron hace algunos años en el centro de atención científica de prestigiosas revistas académicas porque se habían configurado sobre una fórmula desarrollada en los años sesenta en oficinas dominadas por hombres. Como esa fórmula no había sido revisada desde entonces, las mujeres pasaban frío en su trabajo, y estudios recientes denuncian pocos avances en la erradicación de esa fórmula (IM, 2023: 26).

Amazon se ha visto inmersa en un caso con cierta repercusión por establecer un sistema algorítmico para la selección de personal que discriminaba a las mujeres. De nuevo aquí la causa del sesgo estaba en la asimetría de datos, pues el algoritmo había sido entrenado con currículos de candidatos compilados a lo largo de una década y, cuando se trataba de puestos técnicos, la mayoría de esos candidatos eran hombres y la IA infirió que ser hombre era un requisito necesario o conveniente para esos puestos. Los currículos de mujeres fueron sistemáticamente excluidos, a pesar de que contaran con formación técnica adecuada para los puestos ofertados. Amazon intentó corregir el sesgo, pero al no conseguirlo decidió suspender el programa (Rivas, 2020: 230-232).

La cuarta de esas causas diferenciales que propician los sesgos discriminatorios en los sistemas de IA es que los resultados sesgados ofrecidos por un sistema de IA (sea cual sea el origen del sesgo: los datos, el algoritmo, la interfaz o se trate de un sesgo humano) se pueden integrar en la mina de datos que utiliza el propio sistema de IA (sesgos de retroalimentación). Por ejemplo, si un sistema de IA excluye de una beca a una persona joven en atención a circunstancias tales como su código postal (pues las personas con ese código postal, correspondiente a un barrio marginal o racializado o a un medio rural,

tienen más fracasos educativos), aparte de perjudicar a la persona solicitante (lo que se daría también si la decisión de exclusión de la beca fuese humana), confirmaría (y esto lo diferencia de la decisión humana) la aplicación del sesgo para otras personas jóvenes con ese mismo código postal que solicitasen la beca y cuya solicitud fuese evaluada por el mismo sistema de IA u otro semejante. O sea, mientras las decisiones humanas con sesgo discriminatorio consolidan el prejuicio a modo de sumatorio, las decisiones automatizadas lo consolidan en términos de multiplicación.

2. PROBLEMAS DE CALIFICACIÓN JURÍDICA DE LOS SESGOS EN LOS DATOS

¿Cómo se califican conceptualmente todas las anteriores situaciones causadas por los sesgos en los datos a efectos de aplicación de la legislación antidiscriminatoria? La detección de los sesgos discriminatorios derivados de los datos manejados se suele derivar hacia la aplicación del concepto de la discriminación indirecta que exige acreditar una desventaja particular, típicamente en forma de impacto adverso detectable estadísticamente, ofreciendo la posibilidad de una justificación ajena a la causa discriminatoria. Pero la respuesta no resulta tan fácil a causa de varias circunstancias.

En primer lugar, porque la estructura conceptual de la discriminación indirecta se sustenta en la existencia de una disposición, criterio o práctica aparentemente neutros que produce una desventaja particular y que no se justifica en criterios objetivos. Una aplicación desenfadada de este concepto en un contexto de IA sería admitir que la implementación de un sistema de IA es una disposición, criterio o práctica. Sin embargo, tal implementación solo es una forma de realizar una tarea que, si permitida por la ley, es totalmente libre. Por ejemplo, utilizar un sistema de IA en la selección de personal es un método de reclutamiento que, dentro de la ley, una empresa puede elegir libremente (como podría elegir contratar personalmente o hacerlo a través de una consultora de recursos humanos, quien también podría optar por contratar personalmente o con un sistema de IA). En un sistema de IA, la disposición, criterio o práctica se debería buscar en el código fuente, pero este no está siempre al alcance de la persona o empresa usuaria del sistema de IA (quien simplemente puede haber contratado su uso con una empresa que lo despliega, situación bastante habitual), y también parece excesivo exigirle acreditar, para justificar un impacto adverso, la utilización de datos sin sesgos discriminatorios, la trazabilidad de la inferencia estadística a partir de datos no discriminatorios, la utilización de minas de datos no asimétricas y la ausencia de retroalimentación. Además, algunos de esos elementos no necesariamente constituyen una justificación suficiente: en particular la trazabilidad de la

inferencia estadística a partir de datos no discriminatorios no garantiza la ausencia de un resultado discriminatorio, pues los sistemas de IA no distinguen entre correlación y causalidad (por ejemplo, la *start-up* Gild identificó una correlación estadística entre visitar páginas de manga japonés y buenas habilidades para codificar; Ginès, 2024: 8). Si se desconoce el código fuente del sistema de IA, si este procesa con total opacidad los datos disponibles, si realiza inexplicables inferencias estadísticas y si alcanza decisiones automatizadas sin justificaciones claramente articuladas, la conclusión es que «habría que cuestionar esa fácil equiparación de la discriminación algorítmica a un eventual supuesto de discriminación indirecta» (Sáez, 2020: 49). Nada de ello se ajusta a los parámetros tradicionales sobre los cuales se aplica una discriminación indirecta.

En segundo lugar, porque la cuestión se complejiza aún más cuando se trata de la interacción de diversos factores que crea una situación específica de discriminación interseccional (Lousada, 2024). Y es que según el concepto de discriminación interseccional, no basta con ser una mujer racializada, de religión minoritaria o minorizada, mayor, con discapacidad y con orientación no heterosexual, para considerar que esa mujer sufre discriminación interseccional, sino que es necesario que se produzca una situación específica de discriminación por la existencia de un prejuicio que, en el contexto de que se trate, conduce a la exclusión jurídica de esa mujer (con lo cual se estaría discriminando a esa mujer racializada, pero no a un hombre con todas esas circunstancias asociadas, ni a una mujer a la cual le faltase una sola de esas circunstancias). Lo que ocurre es que, dada la amplitud de datos que maneja, la IA puede producir esa exclusión de manera automatizada y sin ningún prejuicio asociado a esa situación. En este contexto digital, resulta muy difícil detectar un prejuicio que constituya la situación específica de discriminación para apreciar una discriminación interseccional y el intento de derivar la situación hacia la discriminación indirecta tropieza con la fragmentación de la muestra a efectos de apreciar un impacto adverso.

3. SESGOS CAUSADOS POR EL PROPIO ALGORITMO: CAUSAS, MANIFESTACIONES Y PROBLEMAS DE CALIFICACIÓN JURÍDICA

Mientras que en los supuestos hasta ahora analizados se ha partido del no acceso al código fuente del sistema de IA, si se ha accedido a tal código fuente se podría verificar que el sesgo está causado por el propio algoritmo, y esto sí nos sitúa, en principio, en un escenario conceptual más transitado desde la perspectiva antidiscriminatoria: quien crea el sistema incluye en el algoritmo una instrucción dirigida a situar a personas, en atención a su sexo u a otro motivo protegido, de manera menos favorable que otras en situación

comparable, extendiéndose la responsabilidad a quien, conociendo esa circunstancia, despliega el sistema de IA (lo que constituiría una discriminación directa); o cuando dicha instrucción, aparentemente neutra, coloca a personas de un sexo o incluidas en un colectivo protegido, en desventaja particular con respecto a otras, salvo que dicha disposición, criterio o práctica puedan justificarse objetivamente en atención a una finalidad legítima y que los medios para alcanzar dicha finalidad sean necesarios y adecuados (lo que constituiría una discriminación indirecta).

Un ejemplo de esta situación fue conocida en 2020 por el Tribunal Ordinario de Bolonia[10]: Deliveroo funciona con un algoritmo de atribución de turnos (llamado Frank) en función de si el *rider* se ha dejado de conectar a turnos con anterioridad. Los sindicatos demandaron porque esa instrucción no distinguía si esa conducta obedecía al ejercicio del derecho de huelga, a una enfermedad o al cuidado de familiares. El tribunal consideró la existencia de discriminación porque, siendo neutra la instrucción, colocó a *riders* en desventaja particular sin evaluación de las razones justificadas de su absentismo (Fernández, 2021; Aragüez, 2024: 149-152). Otro ejemplo similar, también en Italia, lo encontramos en la situación conocida en 2023 por el Tribunal Ordinario de Palermo[11]: la calidad y eficiencia de las personas trabajadoras de la empresa Foodinho (filial italiana de Glovo) se media, entre otros parámetros, por la disponibilidad en horas de alta demanda, lo que perjudicaba a aquellas personas trabajadoras que, por razones familiares, de edad o salud, no podían atender a esas horas.

Pero otras situaciones generadas por el propio algoritmo no son tan sencillas de calificar jurídicamente como discriminación. Los programas de traducción suelen marcar sesgos de género al traducir del inglés, donde predominan nombres neutros, a otros idiomas donde los nombres tienen género, como el francés o el español (Flores, 2023: 99). Por ejemplo, si utilizamos el traductor DeepL Translate (que se publicita como «el mejor traductor del mundo») y, para una traducción del inglés al español, introducimos «the nurse», el resultado en español es «la enfermera» y ofrece como alternativa «el enfermero», mientras que si introducimos otras profesiones, el resultado es el inverso, es decir, «the lawyer» se traduce como «el abogado» y la alternativa es «la abogada». Más llamativo resulta que «the nurse and the doctor» se traduce en español como «la enfermera y el médico» y las alternativas son «el

[10] Sentencia Tribunal Ordinario de Bolonia, Sección Laboral, de 31 de diciembre 2020, causa inscrita con el número 2949/2019.

[11] Sentencia Tribunal Ordinario de Palermo, Sección Laboral y de Previsión, de 17 de noviembre de 2023, causa inscrita con el número 9590/2023.

enfermero y el médico» y «la enfermera y el doctor»; o sea, ninguna alternativa contempla a la médica o a la doctora.

No es fácil subsumir estas situaciones de sesgos de género en los conceptos normativos de discriminación. Como no se puede hablar de impacto adverso, sería una discriminación directa, pero, como el sesgo no obedece a una intención de quien crea o despliega el algoritmo, consistiría en la omisión de no diseñarlo para evitar los sesgos discriminatorios, lo que nos sitúa ante una la discriminación por omisión, categoría hasta hace poco desconocida en la legislación (a ella se refiere la Ley 15/2022, art. 4.1), y que ha suscitado una «atención doctrinal discreta» (García, 2021: 261). De esa incuria surgen varias cuestiones: ¿exige la discriminación por omisión una intervención legislativa para establecer una responsabilidad concreta sobre quien despliega el sistema de IA o bastaría deducirla de la prohibición de discriminación por sexo/género? ¿Podría ampararse quien crea o despliega el sistema en inconvenientes técnicos insalvables para evitar el sesgo en una suerte de excepción de buena fe derivada del estado actual de la técnica que, sin embargo, no encontraría apoyo en las definiciones legales actualmente vigentes sobre buena fe ocupacional o sobre buena fe negocial?

4. SESGOS EN LA INTERFAZ DEL PROGRAMA: CAUSAS, MANIFESTACIONES Y PROBLEMAS DE CALIFICACIÓN JURÍDICA

Según el *Diccionario* de la RAE, la interfaz en informática es la «conexión, física o lógica, entre una computadora y el usuario, un dispositivo periférico o un enlace de comunicaciones». Los sesgos discriminatorios se plantean en particular cuando la conexión es entre una computadora y la persona usuaria. Tales sesgos, como los del algoritmo, también son de diseño del sistema de IA, aunque resultan más perceptibles.

Los ejemplos más destacados son los *chatbots* con voces femeninas que han sido denunciados por la UNESCO (y que, dicho sea de paso, contrasta con el uso de voces masculinas especialmente en publicidad, por resultar de mayor autoridad y otorgar mayor credibilidad; Pérez Ugena, 2024: 314). El informe *I'd Blush If I Could: Closing gender divides in digital skills through education* (2019) pone de manifiesto que en respuesta a la frase «you're a bitch», Siri respondía «I'd blush if I could» (me sonrojaría si pudiera); Alexa, «well, thanks for the feedback» (bueno, gracias por el *feedback*); Cortana, «well, that's not going to get us anywhere» (bueno, eso no nos llevará a ninguna parte); y el asistente de Google, «my apologies, I don't understand» (mis disculpas, no lo entiendo). De esta manera, el diseño de los asistentes virtuales pretende proyectar la idea ficticia de que Siri, Alexa o Cortana, códigos informáticos incorpóreos, son mujeres jóvenes, heterosexuales, serviciales,

tolerantes y receptivas a los intentos sexuales masculinos, e incluso a actos ofensivos (UNESCO, 2019: 20).

En la misma línea de sesgos en la interfaz se encuentra el diseño de los personajes de muchos videojuegos, donde se presentan personajes femeninos hipersexualizados (a veces sin ninguna lógica en relación con el juego: por ejemplo, en juegos de lucha, los personajes masculinos aparecen avituallados de armaduras metálicas propios de la lucha, mientras los personajes femeninos, que también entran en lucha con los demás personajes, aparecen apenas en un bikini impropio para la lucha).

Ambos ejemplos (los *chatbots* con voces femeninas y los personajes femeninos hipersexualizados) serían incardinables en la categoría de discriminaciones directas, y además abiertas (pues se aprecian a simple vista sin mayores dificultades). La cuestión sería, en consecuencia, si se pueden considerar amparadas en una excepción de buena fe negocial sustentada en el estado de la ciencia (que no ha desarrollado todavía técnicas para corregir esos sesgos) o en las preferencias de la clientela (esto muy cuestionable), sin poder desconocer la tendencia general a limitar las excepciones de buena fe que se aprecia tanto en la normativa comunitaria e interna como en la jurisprudencia que la aplica (Lousada, 2014: 191 y ss.; Lousada *et al.*, 2024: 95 y ss.). Pero también es verdad que en algunos casos (como el de las traducciones sesgadas) no están en juego, como en otras ocasiones en que se ha cuestionado la aplicación de la excepción de buena fe, un derecho subjetivo individualizado, sino un interés legítimo más difuso para evitar la desigual valoración de los roles de género evidenciado en el diseño de las interfaces.

Otro sesgo de interfaz en un ámbito diferente, y con un abordaje jurídico asimismo diferente, se ha denunciado en nuestro propio país. La CIVIO (asociación sin ánimo de lucro para promover la transparencia de las instituciones y el acceso a la información pública) denunció que el sistema BOSCO utilizado por la Administración para la asignación del bono eléctrico a las personas vulnerables dejaba fuera a las viudas porque, si entraban por la vía de la pensión de incapacidad permanente o jubilación, se les contestaba «no reúne los requisitos» (aunque son pensionistas de viudedad), y si entraban por la vía del nivel de renta se les contestaba «imposibilidad de comprobar los niveles de renta», dado que reciben una pensión (no una renta). No planteó la CIVIO una denuncia de discriminación (que sería interseccional al afectar a viudas con escasos recursos), sino, a través de la normativa sobre transparencia, una solicitud de acceso al código fuente ante el organismo estatal de transparencia. Se denegó el acceso por motivos de seguridad pública, secreto profesional, propiedad intelectual e industrial, protección de datos personales y riesgo de ataques informáticos si se conocieran las vulnerabilidades del programa. La reclamación ante los Juzgados Centrales de lo Contencioso

Administrativo[12] fue desestimada (Moral, 2022: 487-491), lo que, por cierto, choca con precedentes judiciales francés[13] e italiano[14], en los cuales sí se admitió el derecho de acceso al código fuente atendiendo a la normativa sobre transparencia.

5. VIOLENCIA DE GÉNERO EN EL METAVERSO

En una interfaz inmersiva, como las utilizadas en el metaverso, la persona usuaria entra, a través de una personalidad virtual, en la escena diseñada por el programa informático, viajando y visitando los distintos lugares, observando e incluso interactuando con el entorno virtual y con las personalidades virtuales de otras personas usuarias. El concepto *metaverso*, que junto con el término *avatar* fue acuñado en 1992 por el escritor Neal Stephenson en su novela *Snow crash*, se refiere a «un entorno donde los humanos interactúan social y económicamente como avatares en un ciberespacio», lo que, en lo que ahora nos interesa, se deriva en una «complejidad jurídica […] dado que el mismo hace converger nuestras vidas físicas y digitales […] nos enfrentaremos a nuevos problemas laborales fruto de la disociación virtual que resulta de estas nuevas realidades: espacios virtuales de trabajo, metatrabajadores, metaempresas, procesos de selección virtuales, vigilancia algorítmica […]» (Mercader, 2022: 47-49 *passim*).

Aunque acaso no estemos aún en esta situación, algunos problemas ya se han manifestado en el ámbito de las agresiones sexuales. Una de las primeras denuncias fue la realizada por Jordan Belamire (2016), quien en el juego «QuiVr» denunció haber sufrido una conducta inapropiada por parte de «BigBro442», el cual, identificándola como mujer por su voz en el chat, se acercó a su avatar pellizcando la zona de su pecho y tocándole la entrepierna. Belamire abandonó el juego ante la insistencia de la otra persona, quien hacía caso omiso a sus peticiones de que declinase su actitud. Desde entonces, los medios de comunicación se han hecho eco de otras denuncias similares de agresiones sexuales en el metaverso que, según se está constatando, les producen a las víctimas unos impactos psicológicos similares a los producidos por los abusos reales.

---

12 Sentencia de 30 de diciembre de 2021 del Juzgado Central de lo Contencioso-Administrativo 8, Pto. 18/2019, ECLI:ES:AN:2021:5863.

13 Sentencia de 10 de marzo de 2016 de la 2.ª Cámara, Sección 5.ª, del Tribunal Administrativo de París (n. 1508951/5-2).

14 Sentencia del Consejo de Estado Italiano (Sección VI) número 2270, de 8 de abril de 2019.

Jurídicamente, alguna doctrina penal ya está abordando los problemas de si estas conductas de agresiones sexuales en el metaverso se pueden considerar delictivas. Es verdad que en determinados juegos del metaverso se puede asesinar y ser asesinado, sin que, obviamente, ello constituya un delito; pero también es verdad que ello está en las reglas del juego aceptadas por todas las personas usuarias, y la agresión sexual no. Y, de admitir el carácter delictivo de estas conductas de agresión sexual no consentidas, los problemas jurídicos subsiguientes son si encajarían en alguno de los delitos tipificados en nuestro Código Penal, o sería necesario introducir nuevas tipificaciones para conseguir la represión penal, aparte los problemas (que tampoco son menores) de determinación del *forum delicti comissi*, o sea de competencia judicial (Vasco, 2022).

Mientras no se aclaren todas estas cuestiones, una solución posible, que planteo como opinión personal, sería salirse del metaverso y calificar según el mundo real. Pues bien, en el mundo real el metaverso es un bien o servicio disponible para el público y ofertado por empresas o personas suministradoras, y ello determina su inclusión en el ámbito de aplicación de las Directivas 2000/43/CE (artículo 3.1.h) y 2004/113/CE (artículo 3.1). Si calificamos desde el mundo real conductas como las agresiones sexuales en el metaverso, estamos ante un condicionamiento para que las mujeres puedan acceder a tal servicio, con lo cual son (sin perjuicio de otras consideraciones posibles desde la perspectiva penal) lisa y llanamente discriminaciones sexistas con todas las consecuencias que ello debería conllevar, tanto de la persona usuaria causante de la conducta como la eventual responsabilidad de quienes despliegan los programas que construyen el metaverso por no regular un metaverso libre de violencia de género.

6. SESGOS HUMANOS EN EL ENTORNO DEL SISTEMA DE IA

Hay sesgos causados por conductas del entorno humano del sistema de IA. Así, se habla de sesgos de interpretación cuando el sesgo se deriva de la interpretación, hecha por personas, de los resultados que el sistema propone. Un ejemplo se dio durante la pandemia cuando al analizar, desde una perspectiva de género, la afectación sobre el empleo de la pandemia y las medidas antipandemia se atendía exclusivamente a las personas afectadas por expedientes de regulación de empleo, obviando que las mujeres basan una parte importante de su actividad en el empleo informal. También son sesgos de origen humano los llamados sesgos de contexto, que se pueden producir cuando un sistema programado para un contexto se utiliza en otro diferente contexto.

Ninguno de estos sesgos presenta dificultades específicas de calificación conforme a la normativa antidiscriminatoria. Su eventual interés lo presentan por la posible exigencia de responsabilidades a personas individuales. En todo caso, el riesgo de estos sesgos justifica medidas dirigidas a promover la formación en perspectiva de género y enfoque de derechos humanos de las personas que interaccionan con el sistema de IA. La formación es decisiva en este ámbito, como en otros muchos.

7. SESGOS DE INVISIBILIZACIÓN

A todos los anteriores sesgos se suman los de invisibilización, que, sin crear resultados sesgados, invisibilizan los que hay. Siendo riesgos superpuestos a los de origen, no presentan ni más ni menos problemas de conceptualización del resultado sesgado como discriminatorio, pero sí dificultan su identificación así como su prueba, de ahí la oportunidad de tomarlos en consideración en el análisis de aquellos problemas.

¿Cuáles son las causas de los sesgos de invisibilización? Atendiendo a cómo se configuran objetivamente los sistemas de IA, estos presentan mecanismos específicos de invisibilización por motivos varios, técnicos y jurídicos: la explicación del algoritmo está al alcance de pocas personas técnicas en la materia; son habituales múltiples capas de algoritmos de funcionamiento autónomo, o el código fuente del algoritmo se encuentra protegido por la propiedad intelectual y el secreto empresarial. Por ello, se suele hablar de la caja negra de la IA *(black box)*, que ha sido calificada como un «agujero negro digital» (Rivas, 2020: 159). Incluso hay ciertos sistemas de procesamiento profundo (IA subsimbólica que prioriza el resultado sobre la explicabilidad) que no siempre justifican de manera articulada sus resultados, lo que puede ser necesario para motivar la decisión automatizada adoptada en el seno de una relación jurídica y para fiscalizarla en juicio.

Atendiendo a cómo percibimos subjetivamente los sistemas de IA, tendemos a confiar en la perfección de su funcionamiento (aunque a nivel de conocimiento científico esto ya está superado, no así en el imaginario popular), lo que determina una aceptación acrítica de las decisiones algorítmicas —es el sesgo de automatización, o confianza de las personas en la perfección de las máquinas, actualmente reconvertido en una «fascinación bigdataísta» (Goñi, 2019: 60)—. Pero en el fondo los algoritmos solo son «opiniones incrustadas en las matemáticas», de ahí que el algoritmo acabará sesgado por los prejuicios de quienes lo operan (Aragüez, 2024: 71), y no faltan casos en que se utiliza el algoritmo para «blanquear» decisiones discriminatorias (Sáez, 2020: 45-46).

El fenómeno de invisibilización de la discriminación causado por los sistemas de IA no solo conlleva dificultad para acreditar la propia existencia de

la discriminación; cuando esta se consigue acreditar también «se proyecta en la ausencia de responsabilidad humana sobre los resultados sesgados o discriminatorios» (*ibid.*: 45). «Los sesgos y los responsables de haber adoptado las decisiones correspondientes quedarán ocultados detrás de una fórmula matemática» (Rodríguez, 2024: 15).

Otro aspecto de los sistemas de IA que dificulta tanto la visibilización de la discriminación como la exigencia de responsabilidades es, gracias a la cantidad de datos manejados, la posibilidad de justificaciones alternativas de decisiones aparentemente discriminatorias. Ante la constatación de que las conductoras de Uber cobraban un 30 % menos que sus conductores, en un estudio de economistas de Stanford financiado por la propia Uber se justificó la diferencia según otros datos ofrecidos por el sistema de IA: las mujeres abandonaban el trabajo en promedio a los seis meses y los hombres a los dos años, y ello les impedía aprender mejores rutas u otros trucos de experiencia; las mujeres conducen más lento en promedio, con lo cual ganan menos dinero al final de la jornada; y las franjas mejor retribuidas, como las horas nocturnas o los sábados, son más elegidas por los hombres. Con ello «[se] consigue correlacionar el menor ingreso salarial de las mujeres con múltiples variables que acaban diluyendo toda responsabilidad de la empresa (al menos en apariencia)» (Todolí, 2024: 61-62), y subrayamos eso de *al menos en apariencia*, pues algunas de esas *proxies* pueden incorporar componentes de género que las inhabilitarían como justificaciones válidas.

III. ¿AFECTA EL REGLAMENTO EUROPEO DE INTELIGENCIA ARTIFICIAL A LA NORMATIVA ANTIDISCRIMINATORIA VIGENTE?

Visto lo visto en las páginas anteriores, se habrán podido comprobar los importantes retos que los sistemas de IA plantean desde la perspectiva de la tutela antidiscriminatoria. Podríamos añadir que también los plantean desde la perspectiva más amplia de los derechos humanos, pues también confrontan con la integridad moral, la intimidad, la propia imagen o la protección de datos, en especial los sensibles y más aún los biométricos. Más ampliamente, la implementación de sistemas de IA puede conducir a resultados contrarios a cualquier derecho humano si la dejamos decidir sobre todos los aspectos de nuestras vidas. Ítem más: el despliegue de sistemas de IA susceptibles de influir en la actividad cerebral ha conducido a una nueva categoría de derechos humanos: los neuroderechos (Beltrán, 2023; Garrigues y González, 2024).

Frente a estos retos, las respuestas jurídicas, en líneas muy generales, podrían ser dos diferentes (pero que son perfectamente complementarias): un abordaje *ex post*, reforzando las respuestas jurídicas ante las violaciones efectivas

de los derechos humanos lesionados y en particular frente a las discriminaciones; un abordaje *ex ante*, estableciendo exigencias y controles para el despliegue en el mercado de los sistemas de IA con la finalidad, entre otras, de prevenir esas violaciones, algo similar, salvando las distancias, al control administrativo previo a la comercialización de los medicamentos.

Estados Unidos, China y la Unión Europea (que son los tres grandes gigantes en nuestro mundo globalizado) han abordado la cuestión con mayor o menor regulación, y la regulación no siempre con iguales finalidades. El modelo estadounidense, construido sobre la libertad de mercado, ha conocido una reciente regulación con la Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, del presidente Biden, de 2023[15], «aunque no con la misma intensidad que en la UE» (Presno, 2023a: 127-128). Y el modelo chino, construido sobre el «férreo control de la disidencia» y sistemas de «crédito social que estarían prohibidos en Europa» (*ibid.*: 128-129), también ha conocido una regulación en 2023, tanto de la IA como de la IA generativa, inspirada en los proyectos de la UE y la OCDE, «pero su sujeción […] al desarrollo del modelo de socialismo chino, con las implicaciones que ello supone sobre el respeto a los derechos humanos, evidencian las diferentes concepciones en juego» (Fernández, 2023).

Mientras, la Unión europea ha optado, desde el primer momento, por un sistema regulatorio basado en el enfoque de derechos humanos. Así se plasmó en 2019 en las *Directrices éticas para una IA fiable*, elaboradas por el Grupo Independiente de Expertos de Alto Nivel sobre IA creado por la Comisión Europea[16]. En estas directrices, la IA fiable se define como aquella que a lo largo de todo el ciclo de vida del sistema garantiza su licitud (cumplimiento de la ley), su eticidad (principios éticos) y su robustez (técnica y social). Se enumeran (de manera no exhaustiva) siete principios no vinculantes: agencia y supervisión humana; robustez y seguridad; privacidad y gobernanza del dato; transparencia; diversidad, no discriminación y equidad; bienestar social y ambiental, y rendición de cuentas. Más en concreto en lo que interesa a nuestro estudio sobre sesgos discriminatorios en la IA, el principio de diversidad, no discriminación y equidad se particulariza en una «necesidad de evitar sesgos injustos».

Igualmente, y con intensidad, la integración de la evitación de sesgos injustos está en el corazón del RIA. No es casual que en el preámbulo del RIA

[15] Accesible en la página web oficial de la Casa Blanca: https://is.gd/rkI8J3 (consultado el 01/09/2024).

[16] Accesibles en la página web oficial de la Unión europea: https://is.gd/sw8AzB (consultado el 01/09/2024).

se aluda a los riesgos discriminatorios de la IA: se recuerda la «diversidad, no discriminación y equidad» como uno de los siete principios éticos no vinculantes de las «Directrices éticas para una IA fiable» (apdo. 27) y el derecho a la no discriminación como un valor de la Unión (apdos. 28, 48 y 59), así como la igualdad de las personas con discapacidad (apdo. 80); se destacan los riesgos de discriminación de los sistemas de puntuación ciudadana de las personas físicas (apdo. 31), de identificación biométrica (hasta en tres ocasiones: apdos. 32, 54 y 95) y de detección de emociones (apdo. 44); en el ámbito educativo (apdo. 56), el empleo (apdo. 57), los servicios esenciales (apdo. 58) y el control fronterizo (apdo. 60) se destaca la importancia, para prevenir la discriminación, de los datos de alta calidad (apdos. 67, 68 y 70); y en los modelos de IA de uso general la discriminación se enumera entre los «riesgos sistémicos» (apdo. 110).

Sin embargo, el RIA no es una norma antidiscriminatoria, ni siquiera de derechos humanos, sino que es una norma sobre regulación del mercado interior dirigida a controlar las exigencias para el despliegue en el mismo de los sistemas de IA a través del denominado «sistema de semáforo» (Gamero, 2021: 277 y ss.): semáforo rojo para los sistemas prohibidos (art. 5); semáforo ámbar para los sistemas de alto riesgo (arts. 6 y 7), admitidos siempre bajo ciertas condiciones (arts. 8 a 15) y una regulación detallada (arts. 16 a 49); semáforo verde para los demás sistemas.

Que el RIA no es una norma antidiscriminatoria ni afecta a la normativa antidiscriminatoria de la Unión lo viene a ratificar en su preámbulo (apdo. 45): «El presente Reglamento no debe afectar a las prácticas prohibidas por el Derecho de la Unión, incluido el Derecho de la Unión en materia de […] no discriminación […]».

Y de ahí que, en contraste con su preámbulo, las referencias a la discriminación en el RIA son prácticamente inexistentes (solo dos) en su larguísimo texto articulado.

Una referencia a la no discriminación como un criterio de calidad de los datos de entrenamiento se encuentra en el art. 10.2: dentro de los sistemas de IA de alto riesgo (semáforo ámbar), aquellos que «utilizan técnicas que implican el entrenamiento de modelos de IA con datos se desarrollarán a partir de conjuntos de datos de entrenamiento, validación y prueba que cumplan criterios de calidad», se someterán a «prácticas de gobernanza y gestión de datos adecuadas para la finalidad prevista del sistema de IA de alto riesgo» centrados, entre otras particularidades, en «el examen atendiendo a posibles sesgos que puedan […] dar lugar a algún tipo de discriminación prohibida por el Derecho de la Unión, especialmente cuando las salidas de datos influyan en las informaciones de entrada de futuras operaciones» (letra

f) y en «medidas adecuadas para detectar, prevenir y reducir posibles sesgos detectados» (letra g).

Otra referencia es la atribución (art. 77) a las autoridades encargadas de proteger los derechos fundamentales, «en particular el derecho a la no discriminación», de determinados poderes en relación con sistemas de IA de alto riesgo: solicitar cualquier documentación creada o conservada conforme al RIA, y acceder a ella «en un lenguaje y formato accesibles», así como, en determinadas condiciones, solicitar a la autoridad de vigilancia del mercado «pruebas del sistema de IA de alto riesgo a través de medios técnicos» —un *sandbox* regulatorio, esto es, «espacios controlados de pruebas para la IA establecidos por las autoridades competentes que proporcionarán un entorno controlado que facilite el desarrollo, la prueba y la validación de sistemas innovadores de algoritmos o sistemas de IA durante un periodo limitado antes de su introducción en el mercado o su puesta en servicio, en virtud de un plan específico» (Mercader, 2022: 60)—. La no superación del *sandbox* podría conducir a introducir las correcciones necesarias en el sistema e, incluso, a prescindir del mismo («apagar Skynet»); en alguna ocasión se ha llegado a esta solución por los riesgos apreciados: «Facebook apaga una inteligencia artificial que había inventado su propio idioma» (*El Mundo*, 28/07/2017).

En consecuencia, el derecho de la Unión europea no altera la preexistente normativa antidiscriminatoria, pero sí ha desarrollado un modelo regulatorio basado en la prevención frente a las violaciones de derechos humanos, en particular integrando la tutela antidiscriminatoria dentro del modelo regulatorio implementado. Y esto nos parece muy relevante para la interpretación de la preexistente normativa antidiscriminatoria bajo la hermenéutica de concordancia entre esa normativa y el RIA.

IV. ¿ES NECESARIO UN NUEVO CONCEPTO DE DISCRIMINACIÓN ALGORÍTMICA?

A esta pregunta la doctrina científica ha dado respuestas varias. Para alguna doctrina, «la que podríamos denominar "discriminación algorítmica" se presentará […] como una de las tipologías jurídicas ya existentes, con la única peculiaridad del empleo de algoritmos en su producción» (Preciado, 2021: 17); «no se requiere […] cuerpo jurídico alguno para afrontar viejas formas de discriminación bajo nuevas apariencias como la que proviene del uso de modelos algorítmicos» (Rivas, 2020: 303); «la discriminación algorítmica puede encauzarse en la actual doctrina antidiscriminatoria, sin que sea necesaria la creación de categorías jurídicas nuevas» (Ginès, 2024: 9). Mientras otra doctrina destaca la existencia de especialidades probatorias

significativas (Todolí, 2023: 73 y ss.) o se plantea la necesidad de articular un test particular de discriminación cuando se involucran algoritmos o adaptar el de discriminación indirecta para flexibilizar la prueba o establecer una presunción cuasi objetiva (Pérez, 2023: 188).

Con la finalidad de aportar algunas consideraciones, volveremos a continuación sobre algunos de los problemas de calificación a que se ha hecho alusión a lo largo de la exposición (en el epígrafe 1) bajo la perspectiva (explicitada en el epígrafe 2) de una interpretación concordada entre la normativa antidiscriminatoria preexistente y el RIA.

Vayamos para empezar a lo aparentemente más sencillo: los sesgos discriminatorios en el diseño del sistema de IA (sesgos en el algoritmo o en la interfaz). Y es lo aparentemente más sencillo porque se ajustan bastante bien a los conceptos de discriminación directa e indirecta. Baste con recordar los ejemplos puestos: plataformas de empleo donde se penaliza no acudir a la llamada sin valorar si ello se debió al ejercicio del derecho de huelga, la conciliación familiar, la edad o condiciones de salud; o interfaz de relación con la persona usuaria que, por su configuración, no permite el acceso a determinados colectivos, como el de las mujeres viudas con escasos recursos. Pero ese ajuste bastante bueno a los conceptos de discriminación no excluye matices en su aplicación, en especial de la indirecta, derivados del contexto de la IA.

Aunque el concepto de discriminación indirecta se construye sobre una noción de desventaja particular más amplia que la de impacto adverso, la desventaja particular en el ámbito de la discriminación algorítmica se manifestará siempre como un impacto adverso, lo que nos remite a una constatación estadística usualmente sobre el resultado de las decisiones adoptadas por el sistema de IA (a cuántas mujeres y hombres afecta). También es posible utilizar otros indicios adelantados al resultado: la mala calidad de los datos utilizados en el entrenamiento del sistema de IA, la utilización de *proxies* sospechosos, o una base de datos asimétrica. La utilización de indicios adelantados puede ser útil a los efectos de acreditar una discriminación interseccional, pues a tales efectos atender al resultado puede ser poco convincente por lo exiguo de la muestra.

Frente a ese impacto (que, en terminología procesalista, sería un indicio o principio de prueba de discriminación), la parte demandada puede desmontar el indicio (sería un contraindicio) o acreditar una justificación objetiva, suficientemente probada, de la medida y de su proporcionalidad (sería una prueba plena en contrario), lo que no difiere de la dinámica general de la prueba de la discriminación (Lousada, 2021: 96 y ss.).

La primera posibilidad (contraindicio) puede situarnos ante «batallas de datos estadísticos» (Todolí, 2023: 94) y, para aclarar esas situaciones, sería conveniente una mayor implicación de los organismos de igualdad en la tutela

antidiscriminatoria y los poderes que a estos se les atribuye en el RIA (en su art. 77, analizado *ut supra*) resultarán decisivos. En cuanto a los indicios adelantados al resultado, la parte demandada siempre puede desacreditarlos demostrando la existencia de una corrección humana que «neutraliza el sesgo en la decisión finalmente adoptada» (Ginès, 2024: 12).

La segunda posibilidad (prueba plena en contrario) no difiere especialmente de la dinámica en juicios sobre discriminación, pues la justificación del impacto adverso (recordémoslo) no pivota sobre la implementación de un sistema de IA (que es una forma de realizar una tarea determinada y no una disposición, criterio o práctica), sino sobre las disposiciones, criterios o prácticas inscritas en el código fuente del algoritmo, que serán las mismas que se podrían exigir de no intervenir el algoritmo. Por ejemplo, en el algoritmo se inscribe un criterio de esfuerzo físico para justificar una retribución y ello impacta sobre las trabajadoras, luego la empresa deberá acreditar, para ser absuelta, que ello está justificado en los términos, ciertamente excepcionales, en que se admite.

Hasta aquí, no parece necesarios nuevos conceptos de discriminación, si acaso matices en su aplicación, pero ello es más labor del poder judicial que del legislativo. En consecuencia, en los casos de discriminación en el diseño del sistema de IA sirven los conceptos de la normativa antidiscriminatoria previa sin que el RIA aporte nada al respecto. Ahora bien, la respuesta judicial que, aplicando dicha normativa, aprecie la existencia de discriminación en el diseño del sistema de IA sí debería obligar a revisar los controles que, en su caso, se hubiesen aplicado para autorizar, según el RIA, su despliegue. O sea, el RIA no influye en los conceptos de la normativa antidiscriminatoria, pero la aplicación de estos obliga a revisar los controles del RIA.

Ahora bien, la discriminación algorítmica más paradigmática, y la que plantea mayores problemas de conceptualización según la normativa antidiscriminatoria, es la que se produce por sesgos en los datos (sesgos derivados de la mala calidad de los datos, sesgos por inferencia a partir de una variable *proxy*, sesgos por asimetría en las minas de datos y sesgos de retroalimentación). En principio, su acreditación exige probar un impacto adverso como si fuera una discriminación indirecta, pero, sin embargo, no es posible verificar la justificación sin conocer las disposiciones, criterios o prácticas inscritas en el código fuente del algoritmo, de ahí que, como paso previo, se deberá verificar si la contraparte cumple con las exigencias de transparencia establecidas en el RIA para el despliegue del sistema de IA. Si se han cumplido y se accede al código fuente, entonces podríamos valorar la justificación de las disposiciones, criterios o prácticas inscritas en él en los términos *ut supra* expuestos.

¿Quid iuris si no se han cumplido? Una interpretación concordada entre la normativa antidiscriminatoria y el RIA permite, a nuestro juicio, concluir la existencia de discriminación, siempre que haya impacto adverso y sin posible justificación en contrario, por la falta de transparencia del sistema de IA. En suma, la discriminación algorítmica por los sesgos discriminatorios provenientes de los datos se podría definir, atendiendo a la interpretación concordada expuesta, como el incumplimiento de las exigencias de transparencia establecidas en el RIA para el despliegue del sistema de IA más la acreditación de un impacto adverso, sin admitirse una justificación en contrario.

No se acaban aquí los problemas de conceptualización, desde la perspectiva de la normativa antidiscriminatoria, de los sesgos en los sistemas de IA. También se han apuntado problemas en la discriminación interseccional, por omisión o en relación con la excepción de buena fe que probablemente puedan ser abordados desde una interpretación concorde con el RIA como expresión de la diligencia debida por quienes desplieguen el sistema de IA. En este sentido, y como mero apunte: en relación con la discriminación interseccional, el uso de *sandboxes* regulatorios puede permitir la identificación de situaciones de impacto adverso en donde confluyan varias causas de discriminación; en relación con la discriminación por omisión, el RIA permite marcar un módulo de diligencia debida que no debería ser obviado; y la excepción de buena fe acaso necesite una concreción para no castigar a quienes actúan conforme con el estado de la ciencia o, en su caso, impedir el despliegue si el estado de la ciencia no garantiza la igualdad.

En todo caso, se hace necesaria la reflexión de si debemos esperar a que sean los operadores jurídicos y los tribunales de justicia quienes se enfrenten a estas nuevas problemáticas en la aplicación de los conceptos antidiscriminatorios adaptándolas al contexto de la IA (dándoles una respuesta fundada en derecho que, en relación con las aportaciones expuestas en este artículo doctrinal, acaso sea acorde con las mismas, o no), o si no sería deseable una intervención legislativa dirigida a precisar los conceptos antidiscriminatorios cuando se apliquen en contextos de IA que complemente y coordine con precisión y seguridad jurídica la normativa antidiscriminatoria con el RIA.

Bibliografía

Amoni Reverón, G. A. (2021). Dereitos ante decisións xudiciais e administrativas algorítmicas recoñecidos nos casos CADA, Loomis, Lazio e SyRI. *Revista Administración y Cidadanía. Revista da Escola Galega da Administración Pública*, 16 (2), 125-142. Disponible en: https://doi.org/10.36402/ac.v16i2.4803.

Aragüez Valenzuela, L. (2024). *Hacia la eticidad algorítmica en las relaciones laborales*. Murcia: Laborum.

Belamire, J. (2016). My first virtual reality groping. *Medium*, 20-10-2016. Disponible en: https://clck.ru/re8Jh.

Beltrán de Heredia, I. (2023). *Inteligencia artificial y neuroderechos: la protección del yo inconsciente de la persona*. Pamplona: Aranzadi.

Fernández, C. B. (2023). China aprueba una regulación de la inteligencia artificial y de la inteligencia artificial generativa. *Diariolaley*, 31-08-2023. Disponible en: https://is.gd/ZJ6gXh.

Fernández Sánchez, S. (2021). Frank, el algoritmo consciente de Deliveroo. Comentario a la sentencia del Tribunal de Bolonia 2949/2020, de 31 de diciembre. *Revista de Trabajo y Seguridad Social*, 457, 179-193. Disponible en: https://doi.org/10.51302/rtss.2021.2374.

Flores Anarte, L. (2023). Sesgos de género en la inteligencia artificial: el Estado de derecho frente a la discriminación algorítmica por razón de sexo. *Revista Internacional de Pensamiento Político. 1.ª época*, 18, 97-122. Disponible en: https://doi.org/10.46661/rev.int.pensampolit..8778.

Gamero Casado, E. (2021). El enfoque europeo de inteligencia artificial. *Revista de Derecho Administrativo*, 20, 268-289.

García Campá, S. (2021). Comentario a la Sentencia del Tribunal Superior de Justicia de la Comunidad Valenciana, Sala de lo Contencioso-Administrativo, 134/2021, de 24 de febrero. *Revista de Trabajo y Seguridad Social*, 459, 255-262. Disponible en: https://doi.org/10.51302/rtss.2021.2420.

Garrigues Walker, A. y González de la Garza, L. M. (2024). *Qué son los neuroderechos y cuál es su importancia para la evolución de la naturaleza humana*. Pamplona: Aranzadi.

Ginès Fabrellas, A. (2024). Algoritmos sesgados en el trabajo. Consideraciones en torno a su tratamiento jurídico. *Trabajo y Derecho: Nueva Revista de Actualidad y Relaciones Laborales*, 19, 3.

Giralt García, V. F. (2024). *Prólogo al libro de Lucía Aragüez Valenzuela, Hacia la eticidad algorítmica en las relaciones laborales*. Murcia: Laborum.

Gmyrek, P., Lutz, C. y Newlands, G. (2024). *A Technological Construction of Society: Comparing GPT-4 and Human Respondents for Occupational Evaluation in the UK*. Geneva: International Labour Organization. Disponible en: https://doi.org/10.54394/UQOQ5153.

Goñi Sein, J. L. (2019). Innovaciones tecnológicas, inteligencia artificial y derechos humanos en el trabajo. *Documentación Laboral*, 117, 58-71.

Instituto de las Mujeres (2023). *Informe preliminar con perspectiva interseccional sobre sesgos de género en la Inteligencia Artificial*. Madrid: Instituto de las Mujeres. Disponible en: https://www.inmujeres.gob.es/areasTematicas/SocInfo/Estudios/docs/Informe_Sesgos_Genero_IA.pdf.

Lazcoz Martínez, G. y Castillo Parrilla, J. A. (2020). Valoración algorítmica de los derechos humanos y el Reglamento General de Protección de Datos: el caso SyRI. *Revista Chilena de Derecho y Tecnología*, 9 (1), 207-225. Disponible en: https://doi.org/10.5354/0719-2584.2020.56843.

Lousada Arochena, J. F. (2014). *El derecho fundamental a la igualdad efectiva de mujeres y hombres*. Valencia: Tirant lo Blanch.

Lousada Arochena, J. F. (2021). *La prueba de la discriminación y la lesión de derechos fundamentales (su regulación en los procesos civil, contencioso-administrativo y social)*. Albacete: Editorial Bomarzo.

Lousada Arochena, J. F. (2024). *Mujeres y discriminación interseccional. Un ensayo sobre mujeres en los márgenes.* Madrid: Dykinson. Disponible en: https://doi.org/10.14679/3223.

Lousada Arochena, J. F., Núñez-Cortés Contreras, P. y Cabeza Pereiro, J. (2024). *Igualdad y diversidad en las relaciones laborales*. Valencia: Tirant lo Blanch.

Martínez Garay, L. (coord.) (2024): *Three Predictive Policing Approaches in Spain: VIOGÉN, RISCANVI and VERIPOL: Assessment from a Human Rights Perspective*. Valencia: Publicacions de la Universitat de Valencia. Disponible en: https://is.gd/96ZVOG.

Mercader Uguina, J. R. (2022). *Algoritmos e inteligencia artificial en el derecho digital del trabajo*. Valencia: Tirant lo Blanch.

Moral Soriano, L. (2022). Decisiones automatizadas, derecho administrativo y argumentación jurídica. En F. Llano Alonso (dir.). *Inteligencia artificial y filosofía del derecho* (pp. 475-500). Murcia: Laborum.

O'Neil, C. (2017). *Armas de destrucción matemática. Cómo el big data aumenta la desigualdad y amenaza la democracia*. Madrid: Capitán Swing.

O'Neil, C. (2018). Los algoritmos aumentan las desigualdades sociales. *La Vanguardia*, 04-11-2018. Disponible en: https://is.gd/cO2AQO.

Pérez del Prado, D. (2023). *Derecho, economía y digitalización. El impacto de la inteligencia artificial, los algoritmos y la robótica sobre el empleo y las condiciones de trabajo*. Valencia: Tirant lo Blanch.

Pérez-Ugena Coromina, M. (2024). Sesgo de género (en IA). *Eunomía. Revista en Cultura de la Legalidad*, 26, 311-330.

Preciado Domènech, C. H. (2021). Algoritmos y discriminación en la relación laboral. *Jurisdicción Social*, 223, 5-24.

Presno Linera, M. A. (2022). *Derechos fundamentales e inteligencia artificial.* Madrid: Editorial Marcial Pons. Disponible en: https://doi.org/10.2307/jj.4908196.

Presno Linera, M. A. (2023a). La propuesta de ley de inteligencia artificial europea. *Revista de las Cortes Generales*, 116, 81-133. Disponible en: https://doi.org/10.33426/rcg/2023/116/1775.

Presno Linera, M. A. (2023b). Policía predictiva y prevención de la violencia de género: el sistema VioGén. *Revista de Internet, Derecho y Política*, 39, 1-13. Disponible en: https://doi.org/10.7238/idp.v0i39.416473.

Rivas Vallejo, P. (2020). *La aplicación de la inteligencia artificial al trabajo y su impacto discriminatorio*. Pamplona: Aranzadi.

Rodríguez Fernández, M. L. (2024). Inteligencia artificial, género y trabajo. *Temas Laborales*, 171, 11-39.

Sáez Lara, C. (2020). El algoritmo como protagonista de la relación laboral. Un análisis desde la perspectiva de la prohibición de discriminación. *Temas Laborales*, 155, 41-60.

Soriano, J. B. (2021). Prevalencia y determinantes de la EPOC en España: EPISCAN II. *Archivos de Bronconeumología: Organo Oficial de la Sociedad Española de Neumología y Cirugía Torácica SEPAR y la Asociación Latinoamericana de Tórax (ALAT)*, 57 (1), 61-69. Disponible en: https://doi.org/10.1016/j.arbres.2020.11.010.

Todolí Signes, A. (2024). *Algoritmos productivos y extractives.* Pamplona: Aranzadi.

UNESCO (2019). *I'd Blush If I Could: Closing gender divides in digital skills through education*. Disponible en: https://is.gd/Z6jhkM.

UNESCO (2020). *Artificial Intelligence and gender equality*. Disponible en: https://is.gd/lSNQlb.

Vasco Gómez, A. (2022). El derecho penal español en los delitos sexuales cometidos en el metaverso. Aplicabilidad y soluciones frente a un problema presente y futuro. *Visual Review. Revista Internacional de Cultura Visual*, 1, 2-15. Disponible en: https://doi.org/10.37467/revvisual.v9.3724.