LA IGUALDAD ENTRE HOMBRES Y MUJERES EN EL DERECHO DE LA UE

SEMINARIO PARA ABOGADOS Y JURISTAS
Barcelona, 25 y 26 de abril de 2022

Sesión "IA e igualdad de género"

*Sesgos en el uso de inteligencia artificial para la gestión de las relaciones laborales: análisis desde el derecho antidiscriminatorio de la Unión Europea**

Pilar Rivas Vallejo. Catedrática de Derecho del Trabajo y de la Seguridad Social.
Universidad de Barcelona

pilar.rivas.vallejo@ub.edu

Sumario:. I. INTRODUCCIÓN CONTEXTUAL: CONCEPTOS RELEVANTES. 1. Inteligencia artificial en el derecho. 2. Aprendizaje automático. II. SESGOS DISCRIMINATORIOS EN EL ÁMBITO LABORAL DERIVADOS DEL USO DE IA. 1. Digitalización de viejos prejuicios a la luz de las normas contra la discriminación.2. Impacto de la IA en el ámbito del trabajo asalariado.3. Uso en los procesos de selección y contratación. 4. Aplicaciones destinadas al control y evaluación del trabajo. III. DETECCIÓN DE LA DISCRIMINACIÓN ALGORÍTMICA. 1. Interrelaciones y equivalencias conceptuales entre los lenguajes jurídico y de computación. 2. Insuficiencia del derecho antidiscriminatorio.2.1. Discriminación múltiple e interseccionalidad. 2.2. Disfunciones entre los conceptos de sesgo algorítmico y discriminación por asociación, por error o múltiple. 3. Valoración como discriminatorias de las decisiones automatizadas. 3.1. ¿Es relevante conocer cómo actúa un algoritmo para calificar jurídicamente su impacto como discriminatorio? 3.2. Impacto discriminatorio de los algoritmos. 3.3. Calificación como discriminación directa o indirecta. IV. TUTELA DESDE EL DERECHO ANTIDISCRIMINATORIO. 1. Marcos regulatorios frente a la opacidad de las decisiones automatizadas. 2. Acceso y explicabilidad de algoritmos. 2.1. Derecho de explicabilidad y acceso al razonamiento subyacente. 2.2. Intervención humana significativa. 2.3. Acceso a la motivación y derechos de propiedad intelectual. 3. Indicios y prueba de la discriminación algorítmica. 3.1. Acreditación de los indicios de discriminación en caso de sesgos algorítmicos. 3.2. En caso de discriminación múltiple y/o interseccional. Referencias.

I. INTRODUCCIÓN CONTEXTUAL: CONCEPTOS RELEVANTES

1. Inteligencia artificial en el derecho

La *inteligencia artificial* es el *conjunto de métodos, teorías y técnicas cuya finalidad es reproducir, mediante una máquina, las habilidades cognitivas de los seres humanos* (*Carta ética europea sobre el uso de la Inteligencia Artificial en los sistemas judiciales y su entorno*, de 4 de diciembre de 2018). Para el Consejo Económico y Social Europeo es "la disciplina tendente a utilizar las tecnologías digitales para crear sistemas capaces de reproducir de forma autónoma las funciones cognitivas humanas, incluida la captación de datos y formas de comprensión y adaptación (solución de problemas, razonamiento y aprendizaje automáticos)" (Dictamen *Inteligencia artificial: anticipar su impacto en el trabajo para garantizar una transición justa*, punto 2.2).

Aunque no existe hasta la fecha una definición legal de inteligencia artificial, la propuesta de Reglamento de Inteligencia Artificial de la Unión Europea de 21 de abril de 2021[1], introduce el

*El presente trabajo constituye en parte una versión del que se publicará en el número 1 del año 2022 de E-revista internacional de protección social con el título "Sesgos de género en el uso de inteligencia artificial para la gestión de las relaciones laborales: análisis desde el derecho antidiscriminatorio", *E-revista*

concepto a los efectos de dicha norma, asimilándolo a "software" (a su vez, compuesto de distintas piezas consistentes en algoritmos, Ebers, 2020: 40). En su art. 3 (definiciones) dispone que "se entenderá por 'sistema de inteligencia artificial (sistema de IA)' 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".

La "ciencia de datos" es la disciplina de la computación encargada de procesar macrodatos en todas sus fases, comenzando por su recopilación a través de distintas técnicas (v.g. *data mining* y *reality mining*), y para distintos fines, como los predictivos[2].

El procesamiento de datos (macrodatos) se entiende como "cualquier operación o conjunto de operaciones realizadas sobre datos personales, como la recopilación, el almacenamiento, la conservación, la alteración, la recuperación, la divulgación, la puesta a disposición, el borrado, la destrucción o la realización de operaciones lógicas y/o aritméticas sobre tales datos" (art. 2 b) del Convenio 108, del Consejo de Europa, *para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal*, 1981).

2. Aprendizaje automático

La inteligencia artificial puede observarse desde distintos prismas, como aproximación a la emulación del pensamiento humano, o como sistemas soportados por *software* que, a partir de la interpretación y aprendizaje de macrodatos, simplifican o automatizan determinados procesos a través de algoritmos.

Nos interesa a los efectos analizados la segunda modalidad, a su vez distinguible en dos modelos (los llamados top-down, inspirados en la psicología conductiva y basados en la inductiva mediante el aprendizaje con una muestra de ejemplos y construcción de modelos, y down-top, basada en la neurociencia, representada por el computacionalismo funcional, que no utiliza modelos y en el que se encuentra el aprendizaje profundo, que aprende de sus previas inferencias sobre datos), a veces hibridados (Pujol, 2021: 7), porque es la que sirve para detectar patrones de comportamiento de los que inferir conclusiones en grandes conjuntos de datos. A su vez, sirve para efectuar predicciones de alta precisión aplicables a distintos usos comerciales y empresariales, en definitiva, el *aprendizaje automático* o "conjunto de técnicas que aprenden a encontrar patrones sin instrucciones" ("la rama de la inteligencia artificial relacionada con la función de aprender de la experiencia", Pujol, 2021: 8).

Estos modelos de aprendizaje automático adolecen, sin embargo, de tres importantes inconvenientes:

A) En primer lugar, su diseño calculado: clasificar, ordenar, predecir y procesar macrodatos por un algoritmo tiene un fundamento "político", puesto que puede hacer aparecer la realidad de una determinada manera (Bucher, 2018), y que son susceptibles de manipulación por sus diseñadores.

---

internacional de la protección social. Universidad de Sevilla, 2022, vol. 7, núm. 1, así como del capítulo preliminar de la obra "Discriminación algorítmica en el ámbito laboral: perspectiva de género e intervención" (Rivas Vallejo, P. dir., 2022) y se inscribe en el proyecto de investigación financiado por el Ministerio de Ciencia e Innovación, España: "Discriminación algorítmica: género y trabajo", referencia PGC2018-097057-B-I00.

[1] 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 {SEC(2021) 167 final} - {SWD(2021) 84 final}-{SWD(2021) 85 final}, en https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF.

[2] La datificación permite construir predicciones a partir de la acumulación de datos y su cuantificación, pero también amplía el margen de inexactitud. "Datificar un fenómeno es plasmarlo en un formato cuantificado para que pueda ser tabulado y analizado" (Mayer-Schönberger y Cukier, 2013: 37).

Así, "las decisiones críticas no se toman sobre la base de los datos en sí, sino sobre la base de los datos analizados algorítmicamente" (Pasquale, 2015: 21).

B) En segundo lugar, la arquitectura de datos sobre la que trabajan, igualmente diseñada conforme a unos parámetros "intencionales" en el aprendizaje supervisado, puede estar también condicionada por el tipo y cantidad de datos empleados, pues cabría que estos fueran incompletos, sesgados, en suma insuficientes y no representativos (v.g. cuando trabajan con características personales, con la diversidad y representación suficiente de colectivos, individuos, procedencias, etc.) o mal etiquetados (pues, en definitiva, esta tarea se realiza manualmente por humanos, que también incorporan en el etiquetado sus propios prejuicios)... o mal interpretados (v.g. *sesgo de confirmación*), dando lugar a sesgos algorítmicos.

C) El tercer problema es el de la opacidad (según Burrell –2016–, por tres motivos: secreto corporativo o público intencionado, desconocimiento técnico acerca de su funcionamiento, y, finalmente, el propio diseño del modelo y el desajuste entre la optimización matemática y la interpretación semántica de los datos), a lo que contribuye que sean en buena parte *software* de código cerrado o "sistema de caja negra" (lo que proporciona opciones de solución como el diseño de código abierto o las auditorías de algoritmos), cuya complejidad se basa en la interacción entre datos y código [Burrell, 2016]). Junto a ello, su trascendencia a los efectos jurídicos se basa en que las conclusiones que generan los algoritmos o sistemas de IA tienen una *poderosa legitimidad* (Gillespie, 2016), y, además, influyen indudablemente nuestras decisiones, y las decisiones que otros toman sobre nosotros, promoviendo unas y descartando otras (Mackenzie, 2017, y Laaksonen, Haapoja, Kinnunen y Nelimarkka, 2020).

El aprendizaje automático, en suma, trabaja con correlación de datos, pero no con "causalidad" (Spiegelhalter, 2014), porque es mucho más fácil correlacionar que detectar las causas (Mayer-Schönberger y Cukier, 2013: 5), pues la autonomía del algoritmo (que le permite autoaprender y depurar su funcionamiento) es paralela a su opacidad, ya que dificulta establecer la conexión entre los datos de alimentación y los resultados que ofrece, y, por tanto, el origen de la decisión o elección que proporciona, en definitiva, *el motivo por el que este cree que esa es la mejor elección.* Ello puede generar "falsos positivos" o "falsos negativos" (Flores, Bechtel y Lowenkamp, 2016), en definitiva, predicciones erróneas y, por ende, discriminatorias, como demostró el uso del algoritmo predictor de criminalidad COMPAS en Estados Unidos, difíciles de detectar, en cuanto el algoritmo así entrenado acaba convirtiéndose en una "caja negra" (Slavin, 2011). Asimismo, impide conocer el proceso intermedio para llegar a conclusiones a partir de los *inputs* procesados, porque *el acceso al código fuente del algoritmo no permite conocer realmente el origen de la decisión si el sesgo no está en su diseño sino en su alimentación por datos.* Este elemento cobra singular importancia bajo el prisma del derecho, donde la motivación de las decisiones es fundamental tanto para la justificación de su legalidad como para determinar su validez a la luz de la normativa antidiscriminación.

II. SESGOS DISCRIMINATORIOS EN EL ÁMBITO LABORAL DERIVADOS DEL USO DE IA

1. Digitalización de viejos prejuicios a la luz de las normas contra la discriminación

La falta de transparencia u opacidad propias de los mecanismos automatizados incrementa las habituales dificultades en la reclamación contra las decisiones empresariales, lo que ha generado una creciente preocupación por el tema en el ámbito jurídico[3], aunque no es nueva en el área de la minería de datos o del aprendizaje de máquina y la inteligencia artificial (Hajian, Bonchi y Castillo, 2016). Lo que sí es evidente es la necesidad de abordar su detección y mitigación desde campos ajenos a la

[3] Han sido pioneras las aportaciones de Pasquale y de Barocas-Selbst: Pasquale, F. (2019); Pasquale, F. (2020) y Barocas, S. y Selbst, A. D. (2016). Se recomiendan en particular los trabajos de R. Xenidis citados en otras notas, así como de Zuiderveen Borgesius, F. (2018).

ciencia de la computación (Scantamburlo, 2021: 704), y, en particular, desde el derecho, en cuanto su impacto se traduce en situaciones de discriminación que generan perjuicios hacia las personas en determinados contextos, como es el laboral. Y por cuanto los modelos matemáticos usados para asistir decisiones (automatizadas) pueden encubrir prejuicios sociales así perpetuados (O'Neil, 2017) y proyectar exponencialmente su impacto, especialmente en el mundo del trabajo[4].

Estos sesgos ocultos en ocasiones obedecen a un objetivo claramente discriminatorio (discriminación directa), pero, en la mayoría de los casos, simplemente son provocados por el desinterés hacia su impacto colateral. La hipotética asepsia del algoritmo se presenta como un mecanismo alternativo a los prejuicios o sesgos humanos. Pero, si ese hipotético modelo de objetividad se basa en datos históricos "reales" (captando conductas discriminatorias reales, cuya reiteración en el tiempo detectada por el algoritmo conduzca a que este la identifique como la decisión correcta), su propio diseño escapa del criterio de la objetivación pretendida, sustituido por el de la eficiencia y la productividad. Y, al mismo tiempo, cuando se usa para adoptar decisiones laborales, puede servir para eludir la eficacia del plan de igualdad de las empresas, cuyas medidas se han podido construir precisamente para superar esas situaciones previas que sirven de soporte al modelo automatizado de decisión, lo que implica que, desde la perspectiva de la igualdad entre mujeres y hombres, la progresiva sustitución de mecanismos tradicionales de decisión por estos automatizados asistidos por inteligencia artificial (IA) juega claramente en contra de medidas pactadas para superar las desigualdades relativas al acceso al empleo y a las condiciones de trabajo[5], a menos que el propio plan de igualdad prevea mecanismos de corrección del impacto de herramientas automatizadas, que deberían ser centrales en el diseño de dichos planes.

La aparente neutralidad, objetividad y asepsia de los mecanismos automatizados de decisión juega sin duda en contra de la tutela del derecho a la igualdad, por cuanto la técnica del aprendizaje profundo en la que consisten, a partir de los datos de alimentación por parte de algoritmos que autoaprenden de ellos e infieren conclusiones, utilizadas para asesorar decisiones, impide establecer una clara conexión entre tales datos de alimentación (macrodatos) y las conclusiones a las que llega el modelo matemático. La perversidad de este funcionamiento es, precisamente, la dificultad de conocer dónde se encuentra el sesgo de la decisión si esta es discriminatoria (v.g., en un proceso de selección, por qué ha decidido que un individuo tiene "más valor laboral" que otro), y detectar el error a fin de impugnar la decisión empresarial.

2. Impacto de la IA en el ámbito del trabajo asalariado

La tecnificación de los puestos de trabajo es un fenómeno consustancial a la propia evolución de la tecnología. La inteligencia artificial (IA) constituye un paso más en esta evolución natural y comporta que el desempeño del trabajo también se adapte a las ventajas que ofrece, aunque estas también puedan implicar la introducción de un cierto desorden en la dinámica de las relaciones de trabajo, con la aparición de nuevos riesgos ligados tanto a la ejecución como a la gestión del trabajo.

Pero lo que nos interesa especialmente en este nuevo orden es el elemento más disruptor, que no se halla propiamente en las herramientas implicadas en el desarrollo del trabajo, salvando los entornos digitales o los de trabajo con cobots, sino en la nueva forma de gestionar la relación de trabajo y, especialmente, la selección, evaluación, vigilancia y control de las personas que aspiran a ocupar un puesto de trabajo o que ya han sido contratadas. En efecto, la gestión de los recursos humanos se ha acomodado en buena parte a métodos que facilitan la adopción de decisiones, tanto las de mayor envergadura como las que atienden a la dinámica cotidiana de la gestión empresarial. Se trata de modelos automatizados basados en IA cuyo diseño se ajusta a tales necesidades corrientes de las actividades empresariales, y que orientan a funciones prácticas algunas de las funcionalidades de los

[4] Rosenblat, A. (2018) y Umoja Noble, S. (2018). Asimismo, Angwin, J., Larson, J., Mattu, S. y Kirchner, L. (2016).

[5] Se remite su tratamiento a Rivas Vallejo, P. (2020) y Rivas Vallejo, P., dir. (2022).

modelos matemáticos alimentados con datos masivos para efectuar predicciones o recomendaciones sobre la mejor decisión en el contexto analizado por el modelo a partir de herramientas de aprendizaje automático. Estas ofrecen la ventaja de su incorporación a un software de fácil adquisición y aplicación, que aligeran la cadena de tareas que un algoritmo puede sustituir sin inversión de tiempo alguna, lo que populariza su uso y determina que su aplicación se advierta a corto plazo como masiva y no reservada únicamente a grandes empresas capaces de financiar su diseño y producción o cuyo volumen de gestión haga recomendable el recurso a herramientas para simplificarla, por su utilidad para reducir costos laborales y controlar los flujos de trabajo (Nguyen, 2021: 1), maximizando la eficacia y productividad.

La proliferación del uso de la inteligencia artificial y el crecimiento de la captación de datos masivos ha mejorado las técnicas y ha intensificado su aplicación, cada vez más extendida. Las empresas recurren actualmente a sistemas de información de recursos humanos (Talla, Workplace Analytics... pues precisamente estos procesos, denominados de "back office", son unos de los primeros en automatizarse [Frank, Roehrig y Pring, 2018: 132]), capaces de "adquirir, almacenar, manejar, analizar, recuperar datos y distribuir toda la información computada en pasos anteriores sobre los recursos humanos de una organización" (Pampouktsi, Avdimiotis, Maragoudakis y Avlonitis, 2021).

Su aplicación en el contexto de actividades empresariales ofrece soporte a todo tipo de tareas de gestión, y, entre ellas, la de las relaciones laborales a través de técnicas de analítica de personas (*People Analytics*), subciencia de la minería de datos basada en la inteligencia artificial, para la gestión y evaluación del talento (Bersin, 2019), que sirve para alimentar sistemas de aprendizaje automático a través de los cuales un algoritmo o conjunto de ellos extraen conclusiones e identifican patrones a partir de la observación de la interacción entre datos (masivos). Tales patrones (históricos) permiten efectuar recomendaciones que se ajusten a los mismos, lo que implica replicar decisiones anteriores que pudieran estar condicionadas por prejuicios sociales y, además, convertir este sesgo en modelo para asistir decisiones futuras. Evidentemente, es una herramienta que perjudica a las mujeres en particular, en cuanto los patrones históricos que replican identifican roles de género y prejuicios hacia el trabajo de las mujeres, que históricamente han sido segregadas en el empleo y objeto de peores condiciones de trabajo (las que los actuales planes de igualdad tratan de mitigar a través de medidas de corrección, que, a su vez, podrían desactivar su eficacia mediante la superposición de decisiones automatizadas que buscan precisamente el efecto contrario).

Entre las aplicaciones que en este contexto pueden encontrarse ocupa un papel central la selección y contratación de trabajadores, pero, del mismo modo, los sistemas que permiten evaluar a personas, y, con ello, clasificarlas y priorizarlas, cuentan con la versatilidad necesaria para que tales utilidades sirvan también a otros efectos dentro de la dinámica de una relación de trabajo: promoción profesional, cálculo de retribuciones, permanencia en la empresa y extinción de contratos…

Esta clasificación y valoración de personas puede implicar el tratamiento de datos personales[6], lo que facilita la tutela de las personas implicadas desde el ámbito de protección de los datos personales, pero a menudo los macrodatos usados para la alimentación del sistema (el algoritmo) no tienen tal naturaleza, no ya porque se trate de datos anonimizados, sino porque su procesamiento no se destina a ser utilizados con los propios propietarios de tales datos, sino con terceros, de suerte que los sesgos que pudieran derivarse de las conclusiones a las que llegue tal algoritmo sean aplicables a nuevos sujetos a partir de los datos personales de estos. Ello supone que, dada tal afectación individual, pueda entrar en aplicación el Reglamento 2016/679, del Parlamento Europeo y del Consejo, de 27 de abril de

[6] El procesamiento de datos (macrodatos) se define como "cualquier operación o conjunto de operaciones realizadas sobre datos personales, como la recopilación, el almacenamiento, la conservación, la alteración, la recuperación, la divulgación, la puesta a disposición, el borrado, la destrucción o la realización de operaciones lógicas y/o aritméticas sobre tales datos" (art. 2 b) del Convenio 108, del Consejo de Europa, para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal, 1981.

2016, *relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos* [RGPD]. Si bien su afectación a las decisiones automatizadas y su impacto discriminatorio es muy tangencial en esta norma, la futura Ley de Inteligencia Artificial en la Unión Europea, canalizada a través de la "Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión [COM(2021) 206 final – 2021/106 (COD)]", debería dar respuesta a todas las incógnitas y vacíos actuales de protección que el derecho antidiscriminatorio de la UE solo cubre parcialmente, aunque constituya un marco jurídico suficiente frente a la discriminación por razón de género (no así la discriminación múltiple o interseccional). Precisamente dicha proposición considera sistemas de IA de alto riesgo (art. 6.2 y anexo III, apdos. 1 y 4) los de identificación biométrica y categorización de personas físicas, y *los destinados a la gestión de los trabajadores, empleo y acceso al autoempleo*.

3. Uso en los procesos de selección y contratación

El empleo de robots o algoritmos de selección permite captar, mediante técnicas de psicogenia en la fase de la entrevista digital (llamada "screening"), toda información de interés para la empresa, incluso neutralizando el derecho a mentir en la entrevista de trabajo, blinda la fidelización a la empresa, y abre la posibilidad de incurrir en *profiling o* "aspectismo" basado en el aspecto físico. Y con ello expulsar del mercado de trabajo a individuos a partir de la detección de rasgos psicológicos que los haga "descartables" (v.g. predictibilidad de la tendencia a la depresión, probabilidad de padecer enfermedad mental "no conveniente" para la empresa [Tufekci]; o, entre otros rasgos, de las tendencias sindicalistas de los candidatos contrarias a la cultura de empresa [Gaster, 2020]).

Estas situaciones no son ajenas al mundo de la selección de trabajadores. La novedad reside en la potencialidad discriminatoria de los nuevos sistemas para lograr tales objetivos, basada en dos principales rasgos diferenciales: la opacidad y la confiabilidad por parte de sus usuarios, junto con la capacidad para replicar sesgos anteriores (históricos) y amplificar su impacto. Si se considera que su uso se produce antes de la constitución de la relación de trabajo, la tutela que el derecho dispensa a su potencial discriminatorio es bastante más limitada (así ocurre en el derecho español), aun cuando, bajo el derecho de la UE, estos primitivos estadios de contratación también queden protegidos (cfr. caso Firma Feryn).

Por otra parte, los sistemas automatizados de detección de emociones (psicogenia) también se aplican en sistemas de atención personal basados en bots que están incorporándose al nivel básico de la gestión de personal, y permiten anticipar conductas, lo que otorga una poderosa herramienta de control empresarial hacia los trabajadores.

La aplicación del sistema de perfilación aplicable a las candidaturas en el intercambio de la oferta y la demanda de empleo (como es el caso de los servicios públicos de empleo) se ha convertido, así, en una práctica conocida tanto en el ámbito privado como en el público. Así, en algunos sistemas de empleo europeos se comienza a implantar como método de evaluación de candidatos en el acceso al empleo, como demuestra su implantación en el sistema de empleo austríaco (Allhutter, Cech, Fischer, Grill y Mager, 2020)[7] o la reciente introducción en el sistema público de empleo español (Send@), donde su carácter público no los priva de un posible sesgo, aun mitigado en su diseño.

[7] El resultado de su implantación recibió las críticas de la doctrina comparada, ya que coloca a todos los demandantes de empleo en un ranking, que se salda con una muy baja empleabilidad de las mujeres en función de una serie de variables como es la maternidad, como consecuencia del diseño sesgado del algoritmo sin perspectiva de género (lo que motiva que se evalúe de forma diferente el impacto de la paternidad y cuidado de hijos en hombres que en mujeres), según han señalado Fröhlich, Spiecker y Döhmann (2018). El algoritmo AMS del sistema público de empleo austríaco calcula desde enero de 2019 las

Los algoritmos manejan los datos que caracterizan a los trabajadores potenciales, recuperan información de datos relacionales y no estructurados y formulan un conjunto de recomendaciones (Protasiewicz, Pedrycz, Kozłowski, Dadas, Kopacz y Gałężewska, 2016). Estas operaciones se pueden realizar de manera totalmente automatizada o con la ayuda de asistencia humana, gracias a algoritmos de optimización, heurística, inteligencia artificial, aprendizaje automático, modelos de datos semánticos y sistemas de soporte de decisiones (Protasiewicz, et al., 2016). El aprendizaje automático facilita poner en conexión los datos disponibles desde diversas fuentes –algoritmos de extracción y rastreo de datos– para obtener la mejor elección en la selección de personal –algoritmo de recomendación– bajo la combinación de "datos semánticos" que interactúan entre sí para identificar rasgos personales/profesionales con los que clasificar y evaluar a las personas dentro de un conjunto de individuos (utilizando tres tipos posibles de datos o información: datos estructurados, datos no estructurados, e información proporcionada por los usuarios o solicitantes en este caso, Protasiewicz, et al., 2016).

En el acceso al empleo asistido por tales herramientas, estas técnicas filtran candidaturas, lo que permite automatizar gran parte del proceso de selección, al mismo tiempo que conservan la intervención humana que exige el art. 22 RGPD, al cumplir un papel de herramienta de "asistencia" en el proceso de decisión (sin perjuicio de la capacidad computacional de dirigir por completo tal proceso hasta la emisión y notificación de la decisión adoptada –cfr. robot reclutador Elenius[8] o Amelia de IPSoft–).

4. Aplicaciones destinadas al control y evaluación del trabajo

La simplificación y eficiencia que aportan los sistemas automatizados de decisión conduce a valorar futuras aplicaciones de gestión laboral como la elección de trabajadores afectados por una medida colectiva, el ajuste de la contratación a los periodos punta de producción o de servicios, o la gestión de la prevención de riesgos laborales según predicciones epidemiológicas (como se ha podido constatar con la pandemia del virus SARS-Co-2), entre otras utilidades. No todos ellas son necesariamente negativas, pues podrían explotarse con fines de detección de patrones históricos de discriminación: desde la prevención de riesgos laborales al análisis de la causalidad de la contratación temporal (y la adaptación de la contratación a la predicción de necesidades), pasando por la detección de condiciones más beneficiosas, hasta la extinción de contratos por causas económicas, técnicas, organizativas o de producción.

Los modelos matemáticos para usos relacionados con las relaciones laborales ya venían siendo aplicados en el ámbito de los recursos humanos, aunque su potencial predictivo conoce actualmente una extraordinaria expansión. Y así, estos sistemas se emplean ya en las plataformas digitales de servicios para definir la remuneración de los trabajadores, su reputación digital y sus posibilidades de acceso a las tareas (Comité Económico y Social Europeo, en su Dictamen *Inteligencia artificial: anticipar su impacto en el trabajo para garantizar una transición justa*).

Estas son algunas de las utilidades del aprendizaje automático y otras herramientas basadas en inteligencia artificial con potencial laboral: a) *causalidad de la contratación laboral temporal,* b) *organización del trabajo*, c) *evaluación de trabajadores y valoración del rendimiento o evaluación del desempeño,* d) *cálculo de retribuciones,* e) *prevención de riesgos laborales*, f) *control del trabajo,* g)

probabilidades de empleo de las personas desempleadas inscritas en dicho servicio, clasificándolas en distintos grupos, con diferentes expectativas formativas y de empleo que condicionan los programas públicos de apoyo a la búsqueda de empleo, conforme a distintos criterios que pueden encerrar sesgos como el sexo, la edad, el domicilio o la nacionalidad. Vid. https://www.ams.at/organisation/public-employment-service-austria/working--recruiting---studying.

[8] https://redelenius.com/. Se presenta como un sistema "rápido e incansable", que codifica más de 36.900 millones de datos por segundo, eliminando el riesgo de adquirir prejuicios.

terminación de relaciones de trabajo (OIT, 2021), h) *detección del histórico de discriminación de la empresa*, e i) *otras finalidades*, entre las cuales podría citarse la detección de condiciones más beneficiosas o la negociación del convenio colectivo de empresa..

Una característica común a todos ellos, que centra el interés jurídico, es su opacidad, pues los trabajadores se ven privados de acceso a los criterios de funcionamiento que se les aplican, sin poder conocer que los sistemas de evaluación de rendimiento (cfr. Workplace Analytics) no se basan en criterios inclusivos, sino que pautan tareas y fragmentan minuciosamente los tiempos que dedican a ellas con completa asepsia de factores humanos y entornos que van a tener consecuencias para su posición en la empresa y sus condiciones de trabajo, por condicionar la puntuación que recibirán. Así es como llegamos a situaciones límite, tales como no reservar tiempos para necesidades personales como las básicas de orden fisiológico y otros condicionantes (en la legislación española sí están previstos en el art. 4.4 de la Ley 10/2021, de 9 de julio, de trabajo a distancia, "dentro de la capacidad de actuación empresarial en este ámbito").

Lo cierto es que la razón última de este cambio de paradigma reside en la capacidad de la tecnología de base IA en términos de vigilancia y control de las personas, en este caso los trabajadores. Esta nueva era ha conducido a normalizar la hipervigilancia del trabajo y de los trabajadores ("el jefe constante", Nguyen, 2021), incluso en el ámbito de los tribunales. En España, el Tribunal Supremo admite la validez de la geolocalización de trabajadores (STS núm. 163/2021, de 8 de febrero) dentro de las coordenadas de tiempo y trabajo, con la única condición de no repercutir su coste y mantenimiento sobre los propios trabajadores. Del mismo modo se admite la vigilancia por terceros a través de técnicas de branding, que miden la satisfacción del cliente e incluyen en los mecanismos de reputación empresarial la evaluación del trabajo prestado por sus trabajadores (*begging and bragging*), al mismo tiempo que los propios trabajadores se han convertido en "prosumidores", evaluadores también. Todos ellos acaban participando en la evaluación digital de los prestadores del servicio y, con ello, incidiendo en las condiciones de trabajo o incluso en la permanencia en la empresa de estos… incorporando sus propios sesgos y prejuicios sociales, como parte del engranaje de la gestión de la relación de trabajo…y, finalmente, condicionando también la política de contratación de la empresa en atención a los gustos de los clientes (riesgo proscrito por el derecho de la UE, a tenor de la sentencia Firma-Feryn, C-54/07, de 10 de julio de 2008).

En definitiva, se requiere un nuevo enfoque inclusivo (Maitland, 2019: 150-159), que incluye la supervisión de las métricas empleadas en las distintas aplicaciones usadas en este ámbito, como hiciera el proyecto de ley aprobado por el Senado del Estado de California (Ley AB13 de Responsabilidad de los Sistemas de Decisión Automatizados 2021[9]).

III. DETECCIÓN DE LA DISCRIMINACIÓN ALGORÍTMICA

1. Interrelaciones y equivalencias conceptuales entre los lenguajes jurídico y de computación

Los científicos de la computación y la inteligencia artificial venían refiriéndose a la existencia de sesgos en la inteligencia artificial (IA), que asociaban a diversas tipologías, y que conectaban a un fenómeno holístico donde se integran diversas concepciones, afirmando que, pese a todo, las herramientas

[9] AB 13, 2021-2022 State Assemb., Reg. Ses. (7 de diciembre de 2020), disponible en https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB13. La ciudad de Nueva York ya había promulgado en 2018 la primera ley de responsabilidad algorítmica del país, para supervisar el uso de algoritmos por parte del gobierno, examinar cómo el error y el sesgo entran en su diseño y recomendar medidas que garantizan la precisión y la equidad, (aprobada el 11 de diciembre de 2017 y disponible en https://www.nyclu.org/en/press-releases/city-council -pasa-primer-proyecto-de-ley-nación-dirección-transparencia-sesgo-uso-gubernamental).

basadas en datos siempre serán mucho más exactas que los juicios emitidos por profesionales (Grove et al., 2000).

La incorporación de los juristas a este debate genera una interesante sinergia entre ambos campos del conocimiento, que aporta el imprescindible análisis jurídico del impacto discriminatorio de los algoritmos. Desde esta perspectiva, se ha abordado incluso la traducción del lenguaje computacional al jurídico, cuyas interrelaciones se revelan asimismo como necesarias tanto para la automatización de la aplicación del derecho y de la justicia como para el tratamiento jurídico de los sesgos algorítmicos. Y es que, como subraya Hildebrandt (2018), el lenguaje de computación podría "erosionar la gramática y el alfabeto del derecho positivo moderno", lo que requerirá una nueva hermenéutica que exige la adecuada comprensión del vocabulario y la gramática del aprendizaje automático. En el análisis del impacto de los sesgos, la literatura especializada apunta también a la divergencia de lenguaje (Chouldechova, 2016, sostiene que la noción de discriminación indirecta no es un concepto estadístico en el sentido matemático, sino ético-jurídico, puesto que un instrumento técnico libre de sesgos predictivos puede provocar una discriminación indirecta en función del contexto al que se aplique), y equipara *justicia* con *inexistencia de sesgo*.

Del mismo modo, el concepto de "fairness" o justicia que se emplea en el campo de la computación para el análisis de sesgos (que puede tener una más acotada correspondencia con el de equidad), dista de ajustarse al lenguaje jurídico, toda vez que, en el ámbito del derecho antidiscriminatorio, son otros conceptos los realmente determinantes, como es el caso del equilibrio entre tal impacto y la justificabilidad, proporcionalidad y racionalidad de la decisión, no equiparables propiamente al principio de justicia material en la aplicación del derecho. La Resolución del Parlamento Europeo, de 14 de marzo de 2017, sobre las implicaciones de los macrodatos en los derechos fundamentales: privacidad, protección de datos, no discriminación, seguridad y aplicación de la ley (2016/2225(INI)), alude en su apartado 22 (referido al impacto discriminatorio de los sistemas de algoritmos y conjuntos de datos) a la equidad como criterio que debe guiar el examen de las predicciones basadas en el análisis de datos. Como instrumento de corrección de la ley en su ajuste a las circunstancias concretas del caso específico, introduciendo en su aplicación los criterios informadores de los principios generales del derecho, la equidad implica igualdad e individualización de trato, por lo que obliga a tratar igual a los casos iguales y desigual a los desiguales, combinando la epiqueya aristotélica y la aequitas romana como "atributo ético-jurídico del derecho" (Ruiz-Gallardón, 2017).

En todo caso, es incuestionable la necesidad de aproximar ambos lenguajes (Hildebrandt, 2021: 13), para que la ciencia de la computación incorpore los conceptos jurídicos del contexto para el que se diseña o debe actuar, o, al menos, un lenguaje jurídico común que permita su adaptación a diferentes ámbitos locales o regionales. Pese a ello, existen cuestiones derivadas de la propia complejidad del derecho que dificultan la estricta correspondencia entre los sesgos algorítmicos y la discriminación en sentido jurídico, más centrada en el resultado que en el proceso, donde radica el sesgo algorítmico (Vantin, 2021: 370; Foster, 2004, y Xenidis y Senden, 2020: 172) y, a la vez, "agudizan las tensiones ya existentes en el corpus antidiscriminatorio de la UE" (Xenidis y Senden, 2020: 738).

La injerencia de terceros en el acto de selección en el que se basa la adopción de una decisión, junto con la autonomía del aprendizaje automático, distorsiona tanto la forma en que aquellas se adoptan como el modo de operar los criterios empleados para ello, frente a los mecanismos "humanos" tradicionales, que tampoco se hallan exentos de sesgos ni de arbitrariedad. A estos pueden superponerse las propias indicaciones humanas (v.g. de la empresa) para el diseño del algoritmo, o, en su caso, las de los desarrolladores del software, asumiendo en este caso los sesgos de programación o de entrenamiento que hayan podido intoxicar el modelo si este no se ha entrenado en un contexto diferente a aquel donde deba operar. Pero, ¿pueden asimilarse los sesgos provenientes de mecanismos automatizados basados en IA al concepto jurídico de "discriminación"?

En el ámbito del trabajo, según el Convenio de la OIT sobre la discriminación (empleo y ocupación), 1958 (núm. 111), la discriminación es «cualquier distinción, exclusión o preferencia basada en motivos

de raza, color, sexo, religión, opinión política, ascendencia nacional u origen social que tenga por efecto anular o alterar la igualdad de oportunidades o de trato en el empleo y la ocupación». Si el concepto se corresponde con la diferencia de trato basada en las características personales de un individuo, y no se anuda al propósito de discriminar, sino al resultado (Tomei, 2003), el sesgo -no deliberado- procedente del análisis algorítmico puede identificarse, por tanto, con el concepto jurídico de discriminación, en el contexto del citado convenio. Si se contrasta este concepto con el de las directivas de la Unión Europea contra la discriminación, aplicables en el ámbito del trabajo (Directivas 2000/78, 2000/46 y 2006/54), se confirma que, de igual modo, sea cual sea el ámbito de la directiva, aquel se identifica con la situación en la que "una persona sea, haya sido o pudiera ser tratada de manera menos favorable que otra en situación análoga por alguno de los motivos mencionados en el artículo 1" (de la respectiva directiva: en el caso de la 2000/78, "religión o convicciones, de discapacidad, de edad o de orientación sexual en el ámbito del empleo y la ocupación", en el ámbito de la Directiva 2000/43, raza u origen étnico, y en el caso de la Directiva 2006/54, sexo)[10]. De forma que, sin perjuicio de la delimitación del alcance de la responsabilidad por actos discriminatorios, el concepto de discriminación se identifica con una diferencia de trato basada en una causa protegida[11], *aun cuando esta conducta o acto no sean intencionados o buscado el resultado discriminatorio.*

Este rasgo permite abarcar todos los efectos derivados de sesgos algorítmicos, se trate de sesgos buscados o de sesgos accidentales derivados de la inferencia de datos (incluso los datos de *proxy*, que invalidan la supuesta neutralidad que ofrece la *anonimización* de los datos objeto de tratamiento).

Al mismo tiempo, permitiría cubrir las situaciones interseccionales cuando el rasgo relevante para el algoritmo no sea una categoría protegida, sino otro rasgo "secundario" si aparece o se asocia a un rasgo sí cubierto, salvando la falta de cobertura legal de la interseccionalidad como categoría jurídica propia, en tanto pueda colaborar o influir en el resultado sesgado que provoque sobre categorías sí protegidas (v.g. el rasgo determinante para el algoritmo puede ser el sobrepeso, pero, asociado a una mujer, "categoría" protegida, puede tener relevancia jurídica para informar una potencial reclamación por discriminación basada en el sexo). Lo cual implica conceder un valor jurídico divergente del resultante del sesgo algorítmico, porque, para el sistema de IA, el valor específico estriba en la confluencia de las inferencias (en este ejemplo, sexo + aspecto), mientras que una sola de las causas probablemente habría producido un resultado distinto, que pudiera no ser determinante para justificar un indicio suficiente a fin de plantear una reclamación por discriminación. A título de ejemplo, inferir del código postal o lugar de residencia, o del nombre u otra variable la pertenencia a categorías protegidas implica combinar distintos factores en el análisis, que, por separado, podrían ser "inocuos" o irrelevantes, pero que, actuando conjuntamente, confluyen en un resultado que afecta, en el ejemplo considerado, a personas de cierto origen nacional o étnico, religión o nivel económico, de suerte que esta conclusión sea la que determine el posicionamiento del individuo o individuos en un determinado ranking, orden, o situación, con efectos perjudiciales en el ámbito laboral (v.g. no seleccionado para ocupar un puesto de trabajo). Por otra parte, acreditar el indicio de discriminación se torna especialmente difícil, en tanto que debe consistir en hallar la correlación entre los datos, que puede ser una cuestión de "modelo de caja negra" (propio del aprendizaje profundo, donde se desconocen las razones que motivan el resultado a partir de la introducción de macrodatos [Mayson, 2019]), lo que significa que la IA multiplica el sesgo, pero no contribuye a identificarlo, al basarse, cuando se emplea aprendizaje automático, en correlaciones de datos.

[10] En España, el anteproyecto de Ley para la igualdad real y efectiva de las personas trans y para la garantía de los derechos de las personas LGTBI (https://www.igualdad.gob.es/servicios/participacion/audienciapublica/Documents/APL%20Igualdad%20Trans%20+LGTBI%20v4.pdf) incluye también dichas categorías protegibles, aunque no lo hace dentro de la categoría de "sexo", sino como categoría autónoma.

[11] Art. 21 de la Carta de los Derechos Fundamentales de la Unión Europea: por razón de sexo, raza, color, orígenes étnicos, religión o convicciones.

Los datos son, en consecuencia, el peor obstáculo a la aportación de indicios, sin perjuicio de que siga siendo posible emplear los mecanismos tradicionales al efecto, esto es, la existencia de elementos protegidos en el caso y el resultado negativo para quien los alega. Del mismo modo, el aprendizaje automático permite mayores justificaciones para desvirtuar los indicios (Xenidis y Senden, 2020: 747), toda vez que opera en el terreno de la discriminación indirecta, por tanto, desarticulable por justificación objetiva y razonable. Por otra parte, la víctima será en estos casos menos consciente del sesgo (Ebers, 2020: 79), principal barrera en muchos casos para la visibilización del riesgo y de su impacto.

2. Insuficiencia del derecho antidiscriminatorio

Distintas figuras clave en este campo del derecho de la Unión Europea muestran la distorsión que introduce tratar de encajar la discriminación algorítmica en dicha disciplina en orden a satisfacer la necesaria tutela de tales situaciones. Es el caso del propio concepto de interseccionalidad, o las formas de discriminación por asociación, por error, o la citada discriminación múltiple cuando se tratan de aplicar a decisiones basadas en inferencias de datos realizadas por inteligencia artificial, pues la correlación de características y su interrelación arrojan un resultado no equiparable al que pudieran adoptar humanos, pues en la decisión puede haber primado alguna característica confluyente que pudiera haber pasado desapercibida a estos y que tampoco sepan identificar en el momento de validar la propuesta de decisión que le ofrezca el algoritmo. Pero, sobre todo, porque la identificación de la causa de la discriminación puede permanecer oculta a efectos de activar tales figuras y la tutela que estas brindan. A título de ejemplo, la correlación de características inferidas por un algoritmo puede identificar una relación del individuo evaluado con una condición protegida como base de su decisión de signo peyorativo (discriminación por asociación), pero, al no ser esta directa o perfectamente visible, puede pasar desapercibida en el examen de la detección de causas discriminatorias (v.g. el algoritmo capta la relación del individuo evaluado con una asociación vecinal derivada de su participación en actividades organizadas por esta y, al mismo tiempo, detecta que la mayoría de tales actividades se relacionan con el mundo árabe, y, a continuación, deriva de ello su adscripción a este colectivo, para excluirle del proceso de selección a un puesto de trabajo), o, aún peor, que las características confluyentes en tal análisis no constituyan condición protegida en el derecho de la UE o las legislaciones nacionales pero hayan influido decisivamente en la conclusión adoptada por el sistema de IA.

2.1. Discriminación múltiple e interseccionalidad

La discriminación *múltiple* forma parte de estudios (Comisión Europea, 2007 y 2009), de textos programáticos y de iniciativas legislativas de la Unión Europea, conscientes de la prevalencia del género en este fenómeno multicausal (Serra Cristóbal, 2020: 146), pero no ha sido efectivamente integrada en su *corpus iuris*, lo que impide su efectiva aplicación. No así la discriminación interseccional como forma específica de discriminación donde prevalece el resultado final sobre la suma de causas o multiplicidad de estas (Makkonen, 2002).

La discriminación *interseccional*, término acuñado por Crenshaw (1989), parte de la necesidad de considerar de manera específica las situaciones de discriminación que operan por la confluencia de distintos factores y que provocan un resultado distinto de la mera suma de causas independientes. En el plano digital, el sistema de inferencias que realiza el aprendizaje automático puede rastrear distintos factores, no todos ellos objeto de tutela específica bajo la legalidad de cada país (caso de la obesidad, el aspecto físico, el uso de determinada indumentaria, tatuajes, piercings… fenómeno denominado *profiling* o "aspectismo", actuación basada en el aspecto, que incorpora con frecuencia prejuicios de género), cuya interrelación provoque el resultado final, sea la no priorización en un proceso de selección, sea la evaluación negativa a otros efectos… El análisis interseccional, sostiene Serra Cristóbal

(2020: 157)[12], "permite analizar las interdependencias entre diversos factores de opresión y, de manera simultánea, promover una interpretación indivisible e interdependiente de los derechos humanos".

La falta de positivización hasta ahora de la figura en el derecho de la Unión Europea[13] (donde se alude en los apartados 14 de la Directiva 2000/43 y 3 de la Directiva 2000/78, ambos en el preámbulo y sin contenido concreto anudado en el cuerpo de la norma)[14] ha dejado sin mecanismos jurídicos la tutela efectiva de las situaciones definidas por la confluencia de varias causas de discriminación, como demuestra la STJUE de 24 de noviembre de 2016, asunto C-443/15, *Parris*. De suerte que la interrelación entre los conceptos computacionales y jurídico carece, en la práctica, de efectos concretos, mientras no encuentre respaldo legal, pese a que el impacto del sesgo algorítmico pueda ser mucho mayor, en tanto que las inferencias que puede realizar el aprendizaje automático son capaces de detectar rasgos que a los humanos y sus prejuicios sociales les pasarían inadvertidos, evitando su impacto discriminatorio, lo que determina que esta modalidad de discriminación encierre un potencial mucho más grave que la discriminación simple o monocausal (Xenidis y Senden, 2020: 740), del mismo modo que el aprendizaje automático tiene una más elevada capacidad de generar discriminaciones por asociación, precisamente por su capacidad de inferir correlaciones entre datos (Wachter, 2020: 371). Este potencial discriminatorio se endurece todavía más si se considera su invisibilidad ante las herramientas de control antidiscriminatorias, frente a las cuales puede permanecer oculto o inadvertido. En consecuencia, reorientar el derecho antidiscriminatorio de la UE contribuiría, sin duda, a mitigar los sesgos algorítmicos, lo que intensifica la urgencia de gestionar una cuestión que lleva demasiado tiempo pendiente de resolverse en el ámbito europeo. Una fórmula adecuada para ello es la propuesta por Wachter (2020: 371): la inclusión como categorías jurídicas protegidas de aquellos individuos que se relacionan con las protegidas de una forma clara. En el plano procesal, la admisión plena de la figura reforzaría la tutela para las víctimas de discriminación múltiple e interseccional (también para la discriminación provocada por el uso de sistemas de IA), simplificando la aportación de un solo indicio del sesgo interseccional y facilitando, pues, la actividad probatoria del sesgo algorítmico (Vantin, 2021: 276).

[12] Vid. asimismo Schiek, D. y Lawson, A. (dirs.), 2016, y Serra Cristóbal, R. (coord.), 2013.

[13] También en el español, sin perjuicio de que la Proposición de Ley Integral de Igualdad de 2021 del Grupo Socialista sí la incluya entre sus conceptos en su art. 6.3. discriminación interseccional cuando concurren o interactúan diversas causas de las previstas en esta Ley, generando una forma específica de discriminación" (b). Asimismo, dispone que "en supuestos de discriminación múltiple e interseccional la motivación de la diferencia de trato, en los términos del apartado segundo del artículo 4, debe darse en relación con cada uno de los motivos de discriminación" (en https://www.congreso.es/public_oficiales/L14/CONG/BOCG/B/BOCG-14-B-146-1.PDF). También el proyecto de Ley Orgánica de Garantía Integral de la Libertad Sexual (aprobado por el Consejo de Ministros el 6/7/2021, https://transparencia.gob.es/servicios-buscador/contenido/normaelaboracion.htm?id=NormaEV03L0-20200902&lang=es&fcAct=2021-06-30T12:23:27.739Z) introduce en su art. 2.5 el principio de atención a la discriminación interseccional y múltiple, que define como la superposición a la violencia de otros factores de discriminación,

[14] Para Xenidis y Senden (2020: 738), el concepto de discriminación múltiple sí está presente en la Directiva 2006/54. En el ámbito europeo, la Sentencia del Tribunal Europeo de Derechos Humanos, asunto B.S. vs. España, de 24 de julio de 2012, sí aplica el concepto aludido. Por otra parte, la propuesta de directiva horizontal de 2008, aún no aprobada, 11531/08 SOC 411 JAI 368 MI 246-COM(2008) 426 final (texto consolidado de 2017), alude expresamente a la interseccionalidad y composición de causas de discriminación de entre las protegidas por las directivas, prohibiendo en su art. 2.2 a) la discriminación múltiple.

2.2. Disfunciones entre los conceptos de sesgo algorítmico y discriminación por asociación, por error o múltiple

Las decisiones automatizadas pueden incurrir en discriminación múltiple como consecuencia de su reproducción de la realidad social compleja y donde convergen por intersección distintos prejuicios sociales susceptibles de confluir sobre los mismos individuos (Xenidis, 2021: 739-740).

Identificar y corregir este sesgo cuando proviene de modelos matemáticos para la adopción de decisiones aumenta la complejidad de una tarea *per se* críptica, pues, en términos computacionales, exige la combinación de diversos criterios de corrección para evitar la segregación de una sola de las características a tutelar. Pero, en realidad, la técnica pone al alcance del derecho antidiscriminatorio opciones de mucha mayor complejidad jurídica, pues resulta más accesible diseñar una orden de no discriminar (esto es, de priorizar) determinadas características combinadas que confluyen en un mismo individuo (sexo, origen ético, orientación sexual…) para un algoritmo que responder con las herramientas legales necesarias a un caso de discriminación múltiple. Simplemente porque lo que, para la técnica resulta factible, para el derecho no está aún consolidado, pues el derecho de la UE[15] no ha positivizado aún un concepto legal de discriminación múltiple que permita identificar y tutelar este tipo de intersección discriminatoria, en lugar de mantener su tratamiento como suma de causas independientes (Xenidis, 2021: 741). Por lo que puede concluirse que el derecho de la UE no está preparado para dar respuesta a la discriminación algorítmica, pues permite a distintas manifestaciones del impacto de las decisiones automatizadas escapar de un marco legal encorsetado que adolece de vacíos legales y obliga a constreñir el análisis del ámbito o alcance de la discriminación a unos parámetros simplificados, donde no hay espacio para la discriminación múltiple[16], ni tampoco para la sustitución de criterios asociados a características protegidas por el derecho antidiscriminatorio de la UE (v.g. país de nacimiento como sustituto de origen étnico, como sucedió en la sentencia de 6 de abril de 2017, asunto C-668/15, Jyske Finans A/S) [Gerards y Xanidis, 2020].

Por otra parte, el concepto de *discriminación por asociación* ofrece dudas aun irresolutas. En efecto, en tanto no cabe identificar asociación con correlación de datos de *proxy*, pues este solo es un elemento que detecta características por proximidad, pero que no opera por sí mismo como elemento de discriminación, no resulta posible sostener un concepto "digital" de discriminación por asociación. Pero sí cabe afirmar una clara conexión entre ambos conceptos y un elevado potencial discriminatorio derivado de la mayor precisión que la asociación entre la persona objeto de discriminación y aquella característica presente en otras personas próximas a su círculo que pueda rastrearse por un sistema de IA (v.g. relación con una persona con discapacidad). Pues las correlaciones que pueden ser desconocidas por los humanos (pero que, conocidas, pudieran llevar a discriminar) son más fácilmente captables por un sistema automatizado (por inferencias en conjuntos de datos) y, por ende, pueden amplificar la capacidad asociativa con fines peyorativos (el ejemplo más claro, que proporciona la conocida sentencia Coleman, de 17 de julio de 2008 [asunto C-303/06], es la detección algorítmica de

[15] La STJUE de 24 de noviembre de 2016, asunto C-443/15, *Parris*, rechaza la consideración como discriminación múltiple o interseccional de dos causas si por separado no son discriminatorias, es decir, exige que por separado ambas sean discriminatorias, lo que, en realidad, podría calificarse de superposición de causas, pero no de una interseccionalidad con mayor carga peyorativa asociada a la conjunción de causas en un resultado nuevo. A tenor de la Proposición de Ley Integral de Igualdad promovida por el grupo parlamentario socialista, la discriminación interseccional se define por la concurrencia o interactuación de diversas causas protegidas, generando una forma específica de discriminación (art. 6.3 b). Por otra parte, dicho texto distingue de este concepto el de discriminación múltiple, para asimilarlo al recogido por la STJUE citada, como la situación en la que "una persona es discriminada de manera simultánea o consecutiva por dos o más causas" protegidas (art. 6.3 a).

[16] FRA – Agencia de los Derechos Fundamentales de la Unión Europea: *Inequalities and multiple discrimination in access to and quality of healthcare*, 2010, http://fra.europa.eu/ en/publication/2013/inequalities-discrimination-healthcare.

la conexión con una persona con discapacidad -familiar directo- a través de distintas inferencias, por ejemplo, su afiliación a una asociación de progenitores de personas con diversidad funcional, u otro tipo de datos que evidencien que, pese a que la empresa desconocía este extremo de su vida familiar, existe tal asociación, como rasgo tributario de exclusión en el proceso de selección y contratación o a efectos de otras decisiones empresariales).

En realidad, la correlación de inferencias entre datos implica relacionar entre sí rasgos que el algoritmo o conjunto de algoritmos empleados priorizan o aquellos que descartan, en función del resultado óptimo que proporcione su análisis, lo que acaba implicando que ciertos rasgos o características sean definidos como más adecuados para los fines empresariales pretendidos (v.g. empleo), mientras que otros serán relegados. Sea como sea, en esa operación automatizada será altamente factible detectar rasgos que determinen una discriminación *por asociación*, en el sentido anteriormente indicado.

La *discriminación por error* (aquella que se funda en una apreciación incorrecta acerca de las características de la persona o personas discriminadas) parece identificarse *a priori* con los errores del aprendizaje automático al realizar inferencias que atribuyen a ciertos rasgos consecuencias que pueden no resultar concordantes con la realidad (errores derivados de las reglas de la lógica cuando existe una insuficiente base de datos, sobrerrepresentación o infrarrepresentación de categorías de datos que a su vez están codificando el mundo real). El proceso deductivo mental que conduce a apreciaciones erróneas puede ser humano o computacional, aunque también tiene carácter humano el etiquetado de datos para transformar elementos de la realidad en conocimiento digital, por ser este realizado por humanos que pueden incurrir en errores y sesgos.

Ahora bien, la cuestión principal en este caso es la autoría de tal error, pues, si en el caso humano es claramente atribuible a su incorrecta comprensión de la realidad o de un rasgo concreto de la persona objeto de valoración u observación, en el caso digital este error ha sido cometido por otras personas interpuestas, por tanto de forma indirecta, o bien por el propio mecanismo de aprendizaje automático, en definitiva responsabilidad también de personas. En el primer caso la imputación de responsabilidad es subjetiva, mientras que en el segundo puede ser objetiva si, en el plano jurídico-laboral, se atribuye la responsabilidad a quien emplea las herramientas que conducen al sesgo, si, al fin y al cabo, los empleadores son los sujetos responsables a todos los efectos en su relación jurídica frente a los empleados o "empleables" y se deriva de la regulación de la responsabilidad empresarial en caso de intervención de terceras empresas (subcontratación y transmisión de empresa) o en materia de uso de productos defectuosos en el marco de los riesgos laborales. El modelo de responsabilidad laboral de carácter contractual salva, en estos casos, las dificultades propias de la opacidad y difícil trazabilidad de sujetos responsables en toda la cadena computacional que ha conducido al resultado final, ya sea por el diseño, por la alimentación, por el etiquetado de datos, el entrenamiento en contextos distintos al de su creación, o la aplicación, así como la de identificar a un responsable principal o único (Vantin, 2021: 376).

3. Valoración como discriminatorias de las decisiones automatizadas

3.1. ¿Es relevante conocer cómo actúa un algoritmo para calificar jurídicamente su impacto como discriminatorio?

La pregunta planteada es si es realmente necesario saber cómo ha llegado a una conclusión un sistema de IA, o si debemos atenernos en exclusiva al resultado.

Planteada así, podríamos barajar dos hipótesis de respuesta: a) la *tesis de la conducta*; b) la *tesis del resultado*. Nuestro derecho prima el resultado, pero también concede relevancia a la conducta en sí, porque esta permite determinar el alcance de la gravedad del incumplimiento como de la responsabilidad derivada, aun cuando la mera existencia de un impacto discriminatorio constituye el indicio necesario para desplazar, en un plano procesal, la carga probatoria hacia el autor o autores de la conducta que desencadena el resultado hipotéticamente discriminatorio. En consecuencia, ni es

irrelevante el proceso que conduce al resultado ni los factores que intervienen en este o conocer cómo operan tales sesgos automatizados y cómo calificarlos desde la perspectiva del derecho.

Es evidente, pues, la necesidad de que los expertos en computación e IA dominen la terminología jurídica afectada para tratar no solo de hallar las correspondencias entre ambos lenguajes, sino también para abordar la mitigación de sesgos con el bagaje adecuado, pues las correlaciones de datos no son ni irrelevantes ni neutrales para el derecho.

3.2. Impacto discriminatorio de los algoritmos

A tenor del uso actual de sistemas de IA con afectación de derechos fundamentales, resulta imprescindible su evaluación desde la perspectiva del derecho a la igualdad y a la no discriminación, así como dilucidar si nuestro derecho está preparado para dar respuesta satisfactoria a las reclamaciones sobre responsabilidad por decisiones automatizadas. Ello exige analizar tanto la naturaleza de estas formas discriminación a la luz de las directivas antidiscriminación de la UE (Directivas 2000/78, 2000/43 y 2006/54) como su posible calificación como discriminación directa o indirecta.

El análisis del impacto discriminatorio de las decisiones automatizadas obliga a confrontarlas con las reglas del derecho antidiscriminatorio, y ello lleva a formular diversas cuestiones, no todas ellas conectadas al ámbito de la discriminación: a) ¿puede haber intencionalidad discriminatoria en el uso de un algoritmo que ofrece conclusiones sesgadas?; b) ¿es necesario que la haya o cabe imputar responsabilidad objetiva derivada de la interposición de mecanismos automatizados para adoptar decisiones o bien cabe deducir la nulidad de estas por falta de intervención humana?; c) basar decisiones en técnicas de automatización con escasa intervención humana que provocan sesgos discriminatorios ¿constituye discriminación directa o discriminación indirecta?; d) ¿son suficientes las herramientas de tutela tradicionales o convendría reinterpretar estas formas de discriminación, ya que fueron diseñadas para operar con otros parámetros?

Siguiendo a Miné (2003: 5), “la discriminación directa puede ser intencional y explícita con respecto al motivo prohibido”, “pero, al estar dicha discriminación explícitamente afirmada, en especial en una norma, cada vez con menor frecuencia, el derecho pone el énfasis en el efecto producido por la diferencia de trato, según un concepto objetivo de la discriminación”, y “el carácter intencional de la discriminación ya no constituye un elemento esencial”.

En el ámbito laboral, sigue siendo relevante la diferencia de trato entre dos situaciones comparables, por encima del propósito (económico, discriminatorio o de otro orden). Estas situaciones pueden estar o no condicionadas por el recurso a herramientas técnicas basadas en IA, pero los elementos comparables serán las propias situaciones en sí mismas y no cómo se actuó sobre ellas (en este caso, cómo se tomó la decisión). Así, estos otros aspectos deberán en todo pesar en la valoración del comportamiento hipotéticamente discriminatorio.

La decisión automatizada puede responder tanto al concepto de discriminación directa como al de discriminación indirecta, pues un algoritmo puede estar diseñado con finalidad estricta de descartar ciertas características, por ejemplo, cuando se trata de selección de personal o para realizar una evaluación a partir de criterios aparentemente neutros que perjudiquen a los individuos con ciertas características o que deliberadamente las ignoren cuando son claramente condicionantes de la valoración de su productividad (v.g. enfermedad o discapacidad), lo que podría calificarse como “diseño no inclusivo” del algoritmo. Tanto si se trata de una práctica no neutral como de un uso aparentemente neutro, pero susceptible de implicar una desventaja particular para las personas que respondan a uno o más criterios (o bien supondrían una desventaja particularmente para personas en función del sexo, en relación con las personas del otro sexo), lo cierto es que la decisión empresarial resulta constitutiva de discriminación. A menos que pueda operar la salvedad que desvirtúa la presunción de discriminación indirecta, esto es, que el criterio o práctica sean justificados objetivamente por un objetivo legítimo y los medios usados sean apropiados y necesarios. En

definitiva, el derecho antidiscriminatorio necesita de una reinterpretación a la luz de estas nuevas formas de discriminación tecnológica que permitan ajustar las respuestas legales.

3.3. Calificación como discriminación directa o indirecta

Calibrar el impacto discriminatorio de las decisiones basadas en algoritmos obliga a examinar en primer lugar si realmente la empresa traslada o introduce parámetros de sesgo en el algoritmo cuando adopta tales decisiones. La respuesta no es unívoca, porque el propio diseño del algoritmo se construye sobre la base de órdenes que persiguen un objetivo, y este se define por quienes ordenan su programación, sean estos los empresarios a cuyo uso directo se destina, o sean los comercializadores de software de empresa para su adquisición por empleadores con el fin de organizar sus "recursos humanos". Por tanto, es posible que la respuesta sea negativa (ello constituiría una explícita discriminación directa, insertada en la estructura del algoritmo).

Ahora bien, si un algoritmo de aprendizaje automático funciona como una caja negra, y simplemente valora todas las variables posibles para alcanzar la decisión más acertada, la decisión final (sesgada) podría estar contaminada y constituir un supuesto de discriminación indirecta, pero ajena a las intenciones de los empleadores que lo utilizaron para fundar una decisión. Se plantean varias cuestiones:

a) ¿Es relevante, pues, la propiedad del algoritmo para derivar responsabilidad por discriminación? Considerando la irrelevancia de la intencionalidad para deducir tal calificación como discriminatorio, ¿cuál sería la de usar deliberadamente arquitecturas de sesgo respecto de la mera adquisición de *software* que provoca el mismo resultado no buscado?
b) Si el modelo de funcionamiento del algoritmo empleado, v.g. en la selección de personal, toma como referencia un patrón histórico, esto es, los antecedentes de sesgo de la propia empresa, ¿estaríamos ante un posible caso de *discriminación directa inconsciente*?

La clave seguramente se halla en la actitud de la empresa frente a este riesgo: su pasividad ante el posible efecto perverso de la elección es lo que le acaba convirtiendo en cómplice del algoritmo sesgado. Y, por supuesto, las acciones previas que alimentaron y entrenaron al algoritmo y que este reproduce como modelo ideal a seguir, que fueron en su momento ejecutadas por la empresa, porque en tal caso, si bien no resulta responsable de una orden directa de discriminar (esta sería el diseño ex professo con tal fin), sí es responsable de sus acciones pasadas.

La cuestión es de enorme interés desde la perspectiva del derecho antidiscriminatorio, en tanto permite considerar la responsabilidad por comportamientos pasados cuando estos constituyen la base inconsciente de nuevas decisiones sugeridas por un tercero (el algoritmo) pero asumidas por humanos, como consecuencia de que un algoritmo ajeno a su esfera de decisión ha determinado que deben ser reproducidos para asistir nuevas decisiones, si la empresa desconoce el funcionamiento del aprendizaje automático y sus consecuencias (precisamente por falta de motivación del algoritmo de los elementos en los que se basa para llegar a la conclusión). En este caso, lo verdaderamente relevante será su consentimiento para validar el sesgo que el algoritmo reproduce (confirmar la propuesta del modelo matemático), por lo que, en el plano laboral, el empleador continúa siendo responsable de su acción discriminatoria, e incluso cabe valorar su conducta como constitutiva de discriminación directa y no indirecta (pues la conducta constitutiva de discriminación directa puede basarse en un mecanismo implícito pero consciente, mientras que la discriminación indirecta requiere que el impacto discriminatorio derive de su afectación prioritaria a un colectivo definido por una característica protegida).

De igual modo, el algoritmo puede replicar tanto discriminaciones directas como indirectas pasadas (*patrón histórico*). Y, por ende, el desplazamiento efectivo de las medidas previstas en el plan de igualdad de la empresa para corregir situaciones discriminatorias por razón de género como consecuencia del uso de patrones históricos para tomar decisiones.

El análisis jurídico de la correlación entre los mecanismos reputacionales de empresa en la valoración de trabajadores y las decisiones empresariales tiene, también, una importancia central en esta aproximación, en cuanto es de singular relevancia en la arquitectura de los algoritmos empleados la ausencia de parámetros de orden personal, como la salud, la discapacidad, la conciliación de la vida familiar y laboral, o los derechos sindicales, así como para la corrección del impacto de un sistema de evaluación de estas características. Lo cierto es que los modelos algorítmicos de evaluación del rendimiento están ignorando sistemáticamente criterios de orden jurídico-laboral objeto de especial tutela jurídica frente a la discriminación, como es el caso de la discapacidad o la necesidad de conciliar la vida personal y familiar. La cuestión se puede ejemplificar en la sentencia del Tribunal de Bolonia de 31 diciembre de 2020, núm. 29491, que sitúa este rasgo discriminatorio precisamente en su diseño no inclusivo o huérfano de cualquier factor de corrección de situaciones que servirían para justificar por parte de los repartidores de Glovo la cancelación de su disponibilidad para atender un servicio, regida por el algoritmo *Frank*, lo que motiva que este descabalgue del orden de prioridad de llamamiento a trabajadores que se encuentran en tales circunstancias -legalmente justificadas-, y, consecuentemente, por reiteración en el tiempo, ponga en riesgo incluso la propia pervivencia de su puesto de trabajo.

IV. TUTELA DESDE EL DERECHO ANTIDISCRIMINATORIO

1. Marcos regulatorios frente a la opacidad de las decisiones automatizadas

Si bien las decisiones empresariales generan responsabilidad directa de los empleadores con independencia de cómo se hayan construido o asesorado aquellas, la ocupación del espacio de decisión laboral por algoritmos altera los tradicionales mecanismos de respuesta, al modificar la capacidad defensiva de los trabajadores, porque sus características amparan la supuesta objetividad de tales decisiones y los presentan como opacos e incontestables. En efecto, el problema de la opacidad es una constante en el ámbito de las decisiones, empresariales o públicas (Burrell, 2016). El uso masivo de datos y la convergencia de complejidad y apariencia de objetividad ("turbia ilusión de objetividad" [Stoica, Riederer y Chaintreau, 2018]) plantean un escenario de confianza y diluyen la percepción de opacidad, pese a las consecuencias tangibles de tales decisiones.

Buena parte del problema de la opacidad se centra tanto en la naturaleza inmaterial y técnica de los algoritmos como en su propia configuración jurídica en tanto que se encuentran sujetos a derechos de propiedad intelectual. Ambos elementos permiten a sus usuarios directos mantener una difusa aura de opacidad sobre sus decisiones cuando estas son automatizadas. Frente a esta barrera de la opacidad actúan los mecanismos de tutela que se derivan del Reglamento 2016/679, del Parlamento Europeo y del Consejo, de 27 de abril de 2016, *relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos* [RGPD], y del Convenio 108, para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal, de 28 de enero de 1981[17], ambos relativos a la adopción de decisiones automatizadas, en tanto no se desarrollen instrumentos autónomos en este campo (no lo será tampoco la futura Ley de IA de la UE, iniciada en abril de 2021).

Por tal razón, el marco jurídico aplicable (caracterizado por la centralidad de la protección de datos personales) exige una adecuada combinación entre las normas citadas y el derecho antidiscriminatorio. De suerte que, mientras las primeras permiten delimitar los parámetros de la motivación de las decisiones empresariales, para valorar la validez de tales decisiones de acuerdo con parámetros de transparencia (que incluye su *explicabilidad*, o derecho a una explicación sobre su

[17] Vid. Consejo de Europa (2017): Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, Committee of experts on internet intermediaries, MSI-NET(2016)06 rev6, https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a.

funcionamiento [Goodman y Flaxman, 2017] y sus posibles excepciones, amparadas en derechos de propiedad intelectual y empresarial) y de intervención humana (*ex* art. 22 RGPD), el derecho antidiscriminatorio proporciona las reglas de distribución de la carga probatoria para indagar en la hipotética naturaleza discriminatoria de tales decisiones.

Además de las normas que delimitan las obligaciones relativas a la protección de datos personales y al derecho a la igualdad y no discriminación (de las que se deriva el deber de motivar la decisión automatizada), confluye el marco regulatorio de la propiedad intelectual, que puede introducir complejidad adicional en la determinación de obligaciones y responsabilidades, pues comporta asimismo problemas de foro competente y de imputación o reparto de responsabilidades (Vantin, 2021: 371), por más que en el ámbito laboral tales responsabilidades hayan de dilucidarse siempre al margen de la relación contractual entre empresas y trabajadores, marco delimitador de la respuesta directa de las primeras frente a los segundos. Sin embargo, esta parcelación de la normativa aplicable requeriría, para un mejor ajuste, de un replanteamiento centrado en la protección frente a decisiones automatizadas y, en general, el uso de inteligencia artificial sobre humanos, donde igualmente quedaría comprendida la protección de los datos personales.

Sin perjuicio de las distintas estrategias que pueden ordenarse a la prevención y mitigación del sesgo, como la gobernanza y equidad algorítmica, la evaluación de impacto de riesgos, auditoría de algoritmos y normalización, la supervisión sindical y de la autoridad pública, la mitigación técnica, amén de la transparencia de algoritmos y otras herramientas de intervención estudiadas por los especialistas[18], como las que se pueden enmarcar en la directiva "Wistleblowing" (Directiva (UE) 2019/1937 del Parlamento Europeo y del Consejo de 23 de octubre de 2019, *relativa a la protección de las personas que informen sobre infracciones del derecho de la Unión*, e incluso el uso de los propios datos con el fin de combatir sesgos discriminatorios (equidad algorítmica, Ho y Xiang, 2020)[19], el análisis que sigue se centrará en el ámbito estrictamente jurídico bajo la óptica de la discriminación, en la que se inscribe el presente seminario.

2. Acceso y explicabilidad de algoritmos

Los algoritmos son definidos como "secuencia(s) finita(s) de reglas formales (operaciones e instrucciones lógicas) que hacen posible obtener un resultado a partir de la entrada de información", lo que implica que dos elementos son cruciales desde una perspectiva jurídica: la secuencia de instrucciones (el "código fuente" [20]) y la información o datos que este utiliza (las llamadas "librerías" o conjunto de datos, sobre los que, a su vez, debe proyectarse el análisis jurídico acerca de su propiedad y derechos de uso, según su procedencia, y acceso por cualquier impugnante), sobre los que centrar el llamado derecho de *explicabilidad* y transparencia, equivalente a la motivación de la decisión que ayudan a asistir.

---

[18] Vid., asimismo, Rivas Vallejo (dir.) (2022) para mayores consideraciones al respecto.

[19] En Estados Unidos, la administración Obama (2016) ya planteó tal necesidad, que se ha traducido en la "Guidance for Regulation of Artificial Intelligence Applications" de 17 de noviembre de 2020 (https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf), donde se plantea la "mitigación de sesgos". La Unidad de Disparidad Racial, del gobierno del Reino Unido, recopila, analiza y publica datos gubernamentales sobre las experiencias de personas de diferentes orígenes étnicos con el fin de impulsar cambios de política donde se encuentran disparidades (Centre for Data Ethics and Innovation [Blog], 2020).

[20] El código fuente se puede definir como "el conjunto de líneas de textos, que son las directrices que debe seguir la computadora para realizar dicho programa; por lo que es en el código fuente donde se encuentra escrito el funcionamiento de la computadora", en caracteres alfanuméricos en un lenguaje de programación elegido por programadores (como pueden ser: Basic, C, C++, C#, Java, Perl, Python, PHP). Aplicado a algoritmos, "por código fuente se entiende todo texto legible por un ser humano y redactado en un lenguaje de programación determinado. El objetivo del código fuente es crear normas y disposiciones claras para el ordenador y que este sea capaz de traducirlas a su propio lenguaje" (definición de *Digital Guide Ionos*).

2.1. Derecho de explicabilidad y acceso al razonamiento subyacente

El derecho de la UE conecta ambos núcleos de tutela: los datos personales y las decisiones automatizadas a través de la protección de los datos personales. El puente de conexión hacia la tutela frente a la discriminación es precisamente la afectación de las decisiones automatizadas basadas en datos, en cuanto estas emplean perfilación (*ex* art. 22 RGPD) y pueden contener sesgos, de muy difícil entendimiento, como resultado de la propia complejidad del aprendizaje automático en el que se basan (Gunning, 2017). De ahí que uno de los elementos fundamentales para articular la protección frente a la discriminación algorítmica, hasta alcanzar un mayor desarrollo normativo, sea, precisamente, el análisis del tratamiento de datos, aun admitiendo la limitación del ámbito aplicativo y posibilidades que brinda el art. 22 RGPD y, por supuesto, del art. 25 del mismo texto, que apela a herramientas que se han revelado como absolutamente insuficientes en el campo de las inferencias por aprendizaje automático (la seudonimización y otras técnicas a las que alude dicho precepto para anonimizar datos o despojarlos de rasgos personales, puesto que cualquier tipo de rasgo es susceptible de rastreo e inferencia, lo que supone que tales técnicas no garantizan la igualdad).

En el ámbito europeo, dos instrumentos legales garantizan el derecho a la protección de datos: el Convenio para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal, convenio 108, y el Reglamento (UE) 2016/679 (RGPD).

Por lo que respecta al derecho de transparencia y explicabilidad, puede materializar, en el contexto de una reclamación por discriminación, la justificación objetiva y razonable que resulta exigible a la empresa autora de la decisión cuestionada y que constituye aplicación de la distribución de la carga probatoria en un proceso por discriminación. No obstante, hacer inteligibles los parámetros y criterios de las decisiones que se toman (*principio de transparencia algorítmica*) resulta difícil en estos casos. La explicabilidad de los algoritmos cumple diversas funciones: ayuda a entender su funcionamiento, tanto para diseñadores y desarrolladores como para personas afectadas por sus efectos, también contribuye a la confiabilidad del sistema que los utiliza (y a su auditoría) y, asimismo, permite construir el argumentario jurídico para desvelar su posible ilegalidad (Ebers, 2021: 48), o la de su impacto una vez aplicado en un contexto determinado. Por ello, resulta determinante calibrar el grado de explicabilidad y valorar la exploración neutral de los algoritmos por un tercero que permanezca ajeno a los problemas de propiedad intelectual y secreto empresarial que pueda suponer su revelación a efectos de búsqueda de sesgos (Ebers, 2021: 80). En este ámbito deviene herramienta clave la auditoría de algoritmos.

En la configuración de este derecho, la primera cuestión relevante es la exclusión de la protección a los datos *no personales*, *ex* art. 9.1 b) del Convenio 108 (que garantiza el derecho individual a la comunicación de los datos procesados de una forma inteligible, toda la información disponible sobre su origen, sobre el periodo de preservación, así como cualquier otra información, con el fin de garantizar la transparencia del tratamiento), que prevé una excepción: los datos personales que no se recopilan de los interesados, caso en el que el responsable está exento de la obligación si el procesamiento implica "esfuerzos desproporcionados". Es posible interpretar que el aprendizaje profundo complica especialmente el cumplimiento de este deber, pudiendo identificarse esta situación con los "esfuerzos desproporcionados" a los que se refiere el art. 8.3.

En segundo lugar, la propia regulación del derecho la restringe a la *contratación en línea* o *procesos totalmente automatizados*. En efecto, los interesados tienen derecho a conocer la motivación del algoritmo cuando este se utiliza para elaborar perfiles (el caso regulado por el art. 22 RGPD y cuyo objetivo es cubrir el ámbito público de procesamiento de datos, con fines de orden público, como se desprende de la proposición de Ley de IA de la Unión Europea), es decir, al conocimiento del *razonamiento subyacente* en el procesamiento de datos cuando sus resultados les son aplicados (este es también el sentido apuntado por el dictamen del CESE cuando se refiere a que el principio de transparencia algorítmica consiste en hacer inteligibles los parámetros y criterios de las decisiones que se toman), y a que esta explicación sea proporcionada por humanos. El considerando 71 del

Reglamento predica este derecho de las personas que acceden a *servicios de contratación en red en los que no medie intervención humana alguna,* y mantiene que e*ste tipo de tratamiento incluye la elaboración de perfiles consistente en cualquier forma de tratamiento de los datos personales que evalúe aspectos personales relativos a una persona física* (afiliación sindical, rendimiento en el trabajo, origen étnico o racial, las opiniones políticas, la religión o creencias filosóficas, datos relativos a la salud o datos sobre la vida sexual, o las condenas e infracciones penales o medidas de seguridad conexas, considerando número 75[21]*, en particular para analizar o predecir aspectos relacionados con el rendimiento en el trabajo… en la medida en que produzca efectos jurídicos en él o le afecte significativamente de modo similar.* Literalmente, pues, la norma parece acotar un ámbito restringido del derecho a la explicabilidad, no extensible, por tanto, a decisiones automatizadas en parte del proceso de decisión, aun cuando esta parte sea especialmente relevante en el conjunto de factores que conducen al a decisión final, v.gr. cribado de currículos en un proceso de selección, finalmente "humanizado" al poner al frente de la decisión a humanos, que, como sucedió en el asunto Uber (Tribunal de Ámsterdam, sentencia C/13/692003/HA RK 20-302), puede haberse limitado a convalidar la recomendación del sistema de IA, que, a su vez, es posible que haya empleado un patrón histórico que replique decisiones sesgadas anteriores, o, aunque ello no sea así, pone fin a un procedimiento en el que la fase de cribado de currículos fue íntegramente automatizada, aunque el proceso haya contado con humanos en su fase final. En este ejemplo, si se considera, v.gr., que han sido presentados mil currículos, de los cuales han sido preseleccionados quince para su examen humano, el grueso del sesgo, donde probablemente radique la mayoría de discriminaciones por diversas causas protegidas, habrá quedado incluido en una fase íntegramente automatizada, sobre la cual no se ofrecerá explicación alguna, cuando, como indica el considerando número 71, no se trate de "*servicios de contratación en red en los que no medie intervención humana alguna"*, lo cual puede acontecer porque no sean servicios de contratación en red o porque sí exista alguna intervención humana en la cadena de decisión. De ahí que devenga de singular relevancia la delimitación del concepto, para exigir que se trate de "intervención humana significativa" o "sustancial".

En tercer lugar, el contenido de la explicación se refiere al "razonamiento subyacente", es decir, al propio *mecanismo de razonamiento del algoritmo* (¿el código fuente o su modificación por aprendizaje profundo?), mientras la norma española sólo ampara el derecho a conocer la finalidad del tratamiento, *no cómo este se ejecuta técnicamente*. Lo que suscita un problema de vaciado de contenido en el uso de aprendizaje automático, pues el elemento determinante de la decisión serán los datos. Y, aun cuando los datos sirvan para elaborar perfiles, el alcance del derecho continúa limitándose a esta circunstancia (así como al derecho a ser informado de su derecho a oponerse, de darse las circunstancias del art. 22 del Reglamento, por tanto, también dentro de un marco restringido de aplicación).

Sin embargo, el derecho a la información básica de quien concurre a un proceso de selección y resulta afectado por un algoritmo que toma la decisión a partir de análisis de datos de terceros a tampoco amplía la información que ya obre en su poder, pues aquel ya probablemente conocerá el propósito del tratamiento de los datos y la identidad de sus responsables (que, de no serles conocida, tampoco le va a aportar gran ayuda en la identificación del sesgo sufrido), aunque la norma añade finalmente que *la información básica podrá incluir en estos casos las categorías de datos objeto de tratamiento y las fuentes de su procedencia*. En este supuesto, el análisis de las categorías de datos permitiría acceder al origen del algoritmo decisor, pero de una manera superficial, si no se comparte el *código fuente*, y,

[21] Los "datos sensibles", a tenor del art. 6 del Convenio 108, son categorías especiales de datos protegidos por el citado precepto, que requieren garantías complementarias apropiadas cuando se procesan, especialmente el origen racial o étnico, opiniones políticas, afiliación sindical, creencias religiosas o de otro tipo, vida sexual o salud…de manera autónoma o en combinación con otros datos (*Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data*, 2015).

aun compartiéndolo, puede no resultar tampoco suficiente si no se comparten los datos que alimentan el algoritmo (Huergo Lora, 2020: 54 ss., y Roig, 2020).

2.2. Intervención humana significativa

La intervención humana conectada con las decisiones automatizadas cuenta con un único referente normativo en nuestro derecho positivo: el art. 22 RGPD (y su homólogo en el derecho interno). A tenor de dicho precepto y su interpretación en los considerandos previos, el derecho de explicabilidad queda referido a los *servicios de contratación en red en los que no medie intervención humana alguna*, lo que evidencia una necesidad urgente de ampliar esta estrecha esfera de atención legislativa a la adopción de decisiones laborales automatizadas. Pero, por otra parte, aun admitiendo su interpretación expansiva a otros ámbitos distintos de los servicios de contratación en red, considerando el uso que las empresas realizan de distintas herramientas con soporte de IA para adoptar decisiones (incluida la selección de personal y contratación, no en línea o en red, aunque esta sí sirva para desplegar el proceso de selección o incluso únicamente el de captación de la demanda de empleo, v.gr. cribado de currículos), el concepto de intervención humana debe ser central en su regulación y análisis jurídico. En particular, el concepto de intervención humana (o human-in-command), reivindicado también desde los planos ético y tecnológico, precisa de una acotación normativa que evite soslayar su exigibilidad en contextos laborales. Es decir, que permita excluir convalidaciones automáticas de recomendaciones automatizadas o que implique que la intervención humana aludida sea algo más que procesar la orden de confirmación, y que, por el contrario, tenga carácter substancial o significativo.

Esta idea puede observarse desde dos perspectivas:

a) La de la *intervención humana significativa*. Esta ha de poder explicar que el procesamiento automatizado previo a la decisión humana lo ha sido de aspectos accesorios o en los que no se adopte decisión final definitiva para ningún individuo. Utilizando un símil procedimental, se referiría a los actos que ponen fin a la vía de que se trate, v.g. de acceso a un puesto de trabajo (en lugar de al proceso o al procedimiento).

b) El de la *acotación del concepto de decisión*. En este caso, se trataría de incluir en dicho ámbito de garantías no solo las decisiones que puedan considerarse "finales" o definitivas, v.gr., la decisión de contratación, sino todas aquellas que tengan relevancia para los sujetos implicados, como es la exclusión del proceso de contratación, lo que requeriría que la fase de cribado también fuera supervisada por humanos, quienes, además, asumieran la responsabilidad de los actos de notificación a los interesados, expresando de este modo la *explicabilidad* exigible, y facilitando su impugnación.

En todo caso, podría sostenerse la conveniencia de que ambas vertientes de esta aproximación se incorporaran a los modelos de decisión automatizada en el ámbito del trabajo, porque se trata de decisiones que pueden tener afectación de derechos fundamentales, como el derecho a la igualdad y a la no discriminación (lo que exigiría una hipotética regulación específica de los derechos digitales laborales).

2.3. Acceso a la motivación y derechos de propiedad intelectual

En el marco de la tutela antidiscriminatoria y, en particular, la justificación objetiva y razonable que puede salvar el indicio discriminatorio, en consecuencia, a la motivación de la decisión impugnada y la delimitación del alcance del derecho a una intervención humana (art. 22 RGPD), puede ser esencial determinar hasta dónde es posible acceder a ese instrumento intangible en el que se basó la decisión empresarial (secuenciación en la que consiste el algoritmo), como a los datos que lo alimentaron (librerías o información en forma de macrodatos). La tendencia incipiente, no obstante, es centrar las reclamaciones contra decisiones automatizadas en el acceso a uno solo de tales elementos, al *código fuente*, con el propósito de verificar los defectos de diseño que puedan determinar el perjuicio que sostiene la reclamación (y conocer cómo se adoptó la decisión).

El abordaje jurídico de la cuestión ha de tener en cuenta dos distintas situaciones: a) algoritmos de arquitectura más simple, no basados en aprendizaje automático, donde los datos de alimentación devienen secundarios, como es el caso planteado contra las aplicaciones administrativas para la solicitud de ayudas públicas (en el que el algoritmo se centra en una decisión binaria [sí/no] [SCANTAMBURLO, 2021: 703]), y debe solo determinar si los solicitantes cumplen o no con los criterios previamente introducidos para diseñar el algoritmo); o, b) los basados en aprendizaje automático o alimentación por datos, caso de los algoritmos predictivos usados en la selección de personal, donde los datos son precisamente la clave del aprendizaje del algoritmo para realizar su predicción o selección (v.g. el algoritmo Send@ del Servicio Público de Empleo Estatal español). Mientras en el primer caso seguramente conocer el código fuente o secuencia de programación permitirá adentrarse en el origen de la decisión (o su eventual manipulación sesgada), en el segundo caso, dicho acceso seguramente no aportará la base necesaria para plantear una oposición solvente al perjuicio causado por el sesgo).

Pues bien, las primeras reclamaciones en el plano laboral se han centrado únicamente en el *código fuente* (cfr. trabajadores de de Glovo[22] o de Uber, que solicitaban el acceso al "código fuente" del algoritmo), como forma de garantizar el derecho de transparencia (aunque dicho acceso no garantiza su inteligibilidad, pues precisa de conocimientos técnicos suficientes para su interpretación). Sin embargo, si el *software* patentado o los algoritmos son propiedad intelectual de sus creadores (o adquirentes, si se cedieron sus derechos), es posible que la empleadora usuaria del mismo no ostente titularidad ni poder de disposición alguno (v.g. empresas adquirentes de una licencia de uso) que permita materializar el derecho de transparencia.

Ante una reclamación de este tenor, cabe por la empleadora oponer derechos de propiedad intelectual sobre el algoritmo, que se despliega sobre "las diferentes partes que integran una obra, (…) siempre que contengan determinados elementos que expresen la creación intelectual del autor" (STJUE, Infopaq International, C-5/08, Rec. p. I-6569, apartado 39, y STJUE, Gran Sala, de 2 de mayo de 2012, asunto SAS Institute Inc contra World Programming Ltd., apartado 65). Dicha protección no se extiende sobre las ideas y principios en los que se basan cualquiera de los elementos de un programa de ordenador incluidos los que sirven de fundamento a sus interfaces, de acuerdo con la Directiva 2009/24/CE del Parlamento Europeo y del Consejo, de 23 de abril de 2009, sobre la protección jurídica de programas de ordenador, que, en su considerando undécimo, se refiere explícitamente a los algoritmos: "de acuerdo con este principio de derechos de autor, en la medida en que la lógica, los algoritmos y los lenguajes de programación abarquen ideas y principios, estos últimos no están protegidos con arreglo a la presente Directiva", que deberán protegerse "mediante derechos de autor" en las legislaciones nacionales. Tal protección, conforme al considerando 63 del RGPD, permite ocultar el código fuente en el contexto de una reclamación (del mismo modo, la mera observación, estudio o verificación del funcionamiento de un programa sin autorización previa del titular no constituye una infracción del derecho, conforme al art. 5.3 de la Directiva 2009/24/CE, pero, a tenor de la jurisprudencia de la UE, esta simple observación no es identificable con el acceso al código fuente).

En la misma línea, si la empresa hubiera obtenido una copia con licencia del algoritmo en cuestión, conforme a la Directiva 2009/24/CE, estaría asimismo autorizada a "observar, estudiar o verificar el funcionamiento de un programa de ordenador con el fin de determinar las ideas y los principios implícitos en cualquier elemento del programa", porque estas no están protegidas por los derechos de autor cubiertos por la directiva (STJUE Gran Sala, de 2 de mayo de 2012, asunto SAS Institute Inc

[22] En el primer caso, el Tribunal de Ámsterdam en sentencia C/13/692003/HA RK 20-302 de 11/3/2021 (y C/13/687315/HA RK 20-207 de la misma fecha) resuelve desfavorablemente, en tanto que se proporcionó a los trabajadores una explicación suficiente sobre dicho funcionamiento y esta no fue impugnada por ellos (sin perjuicio de la obligación de proporcionar acceso a los datos personales a los afectados). El segundo fue resuelto por sentencia del Tribunal ordinario de Bolonia de 27/11/2020, que estimó que los algoritmos utilizados por la empresa no eran inclusivos.

contra World Programming Ltd, razonamiento 50). Igualmente el razonamiento 61 de la sentencia identifica esta situación con el mero uso del programa, sin acceso al código fuente, distinguiendo entre ello y estudiar, observar, verificar…, para concluir que "las palabras clave, la sintaxis, los comandos y combinaciones de comandos, las opciones, los valores por defecto y las iteraciones están compuestos por palabras, cifras o conceptos matemáticos que, considerados aisladamente, no constituyen, en cuanto tales, una creación intelectual del autor del programa de ordenador" (apdo. 66), aunque "sólo a través de la elección, la disposición y la combinación de tales palabras, cifras o conceptos matemáticos puede el autor expresar su espíritu creador de manera original y obtener un resultado, el manual de utilización del programa de ordenador, que constituye una creación intelectual (STJUE de 16 de julio de 2009, Infopaq International, C-5/08, Rec. p. I-6569, apdo. 39). En conclusión, aunque la resolución se refiera a otro núcleo de análisis (la copia de un programa informático o parte de él), el tribunal considera que esa combinación a la que llamamos algoritmo sí constituye una creación intelectual, que se documenta y escribe en lenguaje de código, pues "el objeto de la protección conferida por esa Directiva abarca el programa de ordenador en todas sus formas de expresión, que permiten reproducirlo en diferentes lenguajes informáticos, tales como el código fuente y el código objeto" (STJUE de 22 diciembre 2010, *asunto Bezpecnostní softwarová asociase*, apdo. 35).

3. Indicios y prueba de la discriminación algorítmica

3.1. Acreditación de los indicios de discriminación en caso de sesgos algorítmicos

Para que pueda evaluarse el carácter discriminatorio de una decisión, constituye presupuesto necesario la acreditación del indicio de la discriminación, a tenor de las reglas probatorias (art. 8 Directiva 2000/43/CE, art. 10 Directiva 2000/78/CE y art. 19 Directiva 2006/54/CE), aunque resulta difícil responder a la pregunta de si realmente decidir por mecanismos digitales interpuestos dificulta o no la prueba del indicio.

Como hipótesis de partida no cabe rechazar *a priori* que la interpretación de la apariencia de discriminación o indicio suficiente (Carrizosa, 2012: 59-65) -*prima facie*- no quede alterada como consecuencia del uso de un algoritmo, pues estos se presentan precisamente como herramientas para la objetividad y precisión en el asesoramiento de decisiones. En un escenario como el de un proceso de selección de trabajadores, el sesgo del algoritmo, de existir, puede acreditarse por mecanismos probatorios tradicionales (v.gr. comparación entre los individuos elegidos y los excluidos). Ahora bien, admitido el indicio, y considerando la gran precisión con la que puede operar el algoritmo de selección, a menos que el sesgo se encuentre en su propio diseño (código fuente), ¿cómo se efectúa el análisis comparativo en el universo objeto de procesamiento por el mismo frente al individuo excluido de la selección? En otras palabras, si en un proceso de selección con la concurrencia de cientos de candidatos son excluidos grupos de individuos distintos con características diversas protegidas (o incluso no protegidas), la comparación entre la persona reclamante y las no descartadas no permitirá constatar que la causa de exclusión sea únicamente la que esta posee, pues todo un universo de características inferidas habrá podido ser igualmente rechazado).

Respecto del binomio hombres-mujeres, y la sistemática exclusión de las mujeres, el análisis parece sencillo, pero si se trata del rechazo a características correspondientes a distintos grupos (discapacidad, etnia, religión, procedencia, entre otros) y no otros o no su combinación con otros, el examen que corresponde adquiere complejidad, ya que la multitud de variables combinadas por el modelo automatizado permite considerar también que han primado otras características que habrán de identificarse y que podrían pasar desapercibidas en la acreditación del indicio necesario. Resulta aún más complicado calificar como discriminatorias decisiones cuando los criterios empleados por el algoritmo no correlacionan exactamente con características atribuibles a una categoría social protegida, aunque pueden relacionarse igualmente con tal característica. Por ejemplo, tomando como referencia el caso citado por Xenidis (2021: 4), si el fundamento de la decisión del algoritmo se basa en variables como la distancia del lugar de trabajo, esta no determina *per se* la existencia de un criterio

discriminatorio bajo el derecho positivo, pero si tal factor se anuda a una procedencia concreta, y esta a una característica protegida, la inferencia realizada por el algoritmo deriva en una decisión discriminatoria (v.g. los individuos que residen en la zona rechazada son mayoritariamente población inmigrante). Si, a su vez, la característica con la que se realiza la asociación no se encuentra entre las protegidas por el derecho antidiscriminatorio, se podría producir una situación de discriminación interseccional producto de la confluencia de varios factores que solo un análisis más exhaustivo podría visibilizar y que únicamente un marco de protección más amplio que el proporcionado por las directivas podría calificarse de discriminatorio. Ahora bien, lo que también es importante tener en cuenta es la posible ruptura del nexo de conexión necesario para establecer la preceptiva relación entre la decisión y la discriminación pretendida, precisamente por la distancia entre los datos de referencia y la consecuencia analizada. En definitiva, lo más relevante es que la propia dinámica de funcionamiento de los mecanismos de decisión automatizada dificulta sobremanera la detección de sesgos discriminatorios, y obliga en consecuencia a perfeccionar la granularidad del análisis y reconsiderar la estrategia jurídica frente a la tutela de la discriminación.

3.2. En caso de discriminación múltiple y/o interseccional

La discriminación múltiple e interseccional encuentran una directa equivalencia entre los métodos de inferencia de datos y el resultado discriminatorio, lo que significa que la admisión legal de la autonomía de ambas figuras permitiría dar justa cobertura a bolsas de discriminación que, por acción de los estudiados mecanismos automatizados, no solo pueden estar siendo objeto de un incremento exponencial en buena parte invisible, sino que, además, permanecen al margen de la adecuada tutela jurídica.

Descendiendo al plano probatorio, resulta de interés analizar si la equivalencia entre la interseccionalidad y el sesgo resultante de la inferencia entre campos de datos es susceptible de acreditación probatoria. En la medida en que el aprendizaje automático no permite conocer la correlación entre datos y cuáles de ellos han determinado el resultado, es decir, no es posible saber cuáles de los rasgos analizados han sido determinantes para el resultado ofrecido por el mecanismo automatizado, y, si técnicamente no es factible asistir esta explicación a la hipotética víctima de la decisión discriminatoria, la consecuencia es que probablemente estemos ante un caso de discriminación múltiple, interseccional, pero no sea posible constatarlo, salvo que la empleadora proporcione también a los datos de contraste (no los datos de alimentación). En definitiva, se trata del esquema tradicional de aportación de indicios, facilitado por el uso de otro sistema automatizado capaz de hallar la correlación entre un conjunto de individuos y el individuo que alega la discriminación, que bien pudieran también ser asistidos por mecanismos automatizados capaces de detectar la incidencia estadística simple o interseccional de sesgos discriminatorios.

Referencias

ALLHUTTER, Doris, CECH, Florian, FISCHER, Fabian, GRILL, Gabriel y MAGER, Astrid (2020): "Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective". *Front big data*. DOI: 10.3389/fdata.2020.00005.

ARAGÜEZ VALENZUELA, Lucía (2021): "Los algoritmos digitales en el trabajo. Brechas y sesgos". *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo*. Volumen 9, número 4. ADAPT University Press.

BAROCAS, Solon y SELBST, Andrew D. (2016): "Big data's disparate impact". *California Law Review*, núm. 104, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

BERSIN, Josh (2019): "The Skills Of The Future Are Now Clear: And Despite What You Think, They're Not Technical". Blog del autor, https://joshbersin.com/2019/09/the-skills-of-the-future-are-now-clear-and-despite-what-you-think-theyre-not-technical/ (8/9/2019).

BUCHER, Taina (2018): *If... Then: Algorithmic Power and Politics*. Oxford: Oxford University Press, DOI: 10.1093 / oso / 9780190493028.001.0001.

BURRELL, Jenna (2016): "How the machine 'thinks': understanding opacity in machine learning algorithms". Vol. 3, núm. 1, https://doi.org/10.1177/2053951715622512.

CARRIZOSA, Esther (2012): "La concreción de los indicios de discriminación en la jurisprudencia comunitaria: STJUE 19 abril 2012", *Aranzadi Social*, vol. 5, núm. 7, pp. 59-65.

CHOULDECHOVA, Alexandra (2016): "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments", pp. 1-17, en https://arxiv.org/abs/1610.07524.

CRENSHAW, Kimberle (1989): "Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics". *The University of Chicago Legal Forum: 1989*. HeinOnline - 1989 U. Chi. Legal F. 139, https://philpapers.org/archive/CREDTI.pdf?ncid=txtlnkusaolp00000603.

EBERS, Martin (2020): "Ethical and legal challenges", en EBERS, Martin, y NAVAS, Susana, dirs. (2020): *Algorithms and law*. Cambridge University Press.

FLORES, Anthony W., BECHTEL, Kristin, y LOWENKAMP, Christopher T. (2016): "False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias': There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks". *Federal Probation*. Volumen 80, núm. 2.

GERARDS, Janneke y XENIDIS, Raphaële (2020): "Algorithmic discrimination in Europe: Challenges and Opportunities for EU equality law", *European Futures*, 3/12/2020, https://www.europeanfutures.ed.ac.uk/algorithmic-discrimination-in-europe-challenges-and-opportunities-for-eu-equality-law/.

GILLESPIE, Tarleton (2016): "Algorithm". En *Digital Keywords: A Vocabulary of Information Society and Culture*, ed B. Peters (Princeton, NJ: Princeton University Press), doi: 10.1515/9781400880553-004, y https://www.researchgate.net/publication/309964434_2_Algorithm.

GROVE, William M., ZALD, David H., LEBOW, Boyd S., SNITZ, Beth E. y NELSON, Chad (2000): "Clinical versus mechanical prediction: a meta-analysis". *Psychological assessment*, vol. 12, núm. 1.

GUNNING, David (2017): "Explainable Artificial Intelligence (XAI)", https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf.

HAJIAN, Sara, BONCHI, Francesco, y CASTILLO, Carlos (2016): "Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining", DOI: 10.1145/2939672.2945386, https://www.researchgate.net/publication/305997939_Algorithmic_Bias_From_Discrimination_Discovery_to_Fairness-aware_Data_Mining.

HILDEBRANDT, Mireille (2018): "Algorithmic regulation and the rule of law". *Philosophical Transactions of the Royal Society A*, vol. 376, núm. 2128. DOI:http://dx.doi.org/10.1098/rsta.2017.0355.

HILDEBRANDT, Mireille (2021): "The issue of bias. The framing powers of ML". *Computer Science*, DOI:10.2139/ssrn.3497597. En M. Pelillo, T. Scantamburlo (eds.): *Machine We Trust. Perspectives on Dependable AI*, MIT Press 2021, http://dx.doi.org/10.2139/ssrn.3497597. Preprint.

HO, Daniel E., y XIANG, Alice (2020): "Affirmative Algorithms: The Legal Grounds for Fairness as Awareness". *The University of Chicago Law Review Online* (30/10/2020), https://lawreviewblog.uchicago.edu/2020/10/30/aa-ho-xiang/.

HUERGO LORA, Alejandro (2020): "Una aproximación a los algoritmos desde el Derecho administrativo", en HUERGO LORA, Alejandro (dir.) y DÍAZ GONZÁLEZ, Gustavo Manuel: *La regulación de los algoritmos*. Aranzadi, Cizur Menor.

LAAKSONEN, Salla-Maaria, HAAPOJA, Jesse, KINNUNEN, Teemu, NELIMARKKA, Matti, y PÖYHTÄRI, Reeta (2020): "The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring". *Front. Big Data*, 5/2/2020, https://doi.org/10.3389/fdata.2020.00003.

MACKENZIE, Adrian (2017): *Machine Learners: Archaeology of Data Practice*. Cambridge, MA: The MIT Press.

MAKKONEN, Timo (2002): Multiple, Compound and Iintersectional Discrimination: bringing the experiences of the most marginalized to the fore. Institute For Human Rights, Abo Akademi University.

MAYER-SCHÖNBERGER, Viktor y CUKIER, Kenneth (2013): *Big Data.* Turner publicaciones, Madrid. En http://catedradatos.com.ar/media/3.-Big-data.-La-revolucion-de-los-datos-masivos-Noema-Spanish-Edition-Viktor-Mayer-Schonberger-Kenneth-Cukier.pdf.

MAYSON, Sandra G. (2019): "Bias In, Bias Out". The Yale Law Journal, vol. 128, núm. 8, https://www.yalelawjournal.org/article/bias-in-bias-out#:~:text=abstract.,to%20have%20disparate%20racial%20impacts.

MINÉ, Michel (2003): "Los conceptos de discriminación directa e indirecta", Conferencia "Lucha contra la discriminación: Las nuevas directivas de 2000 sobre la igualdad de trato", 31/3-1/4/2003, Trier, http://www.era-comm.eu/oldoku/Adiskri/02_Key_concepts/2003_Mine_ES.pdf.

NGUYEN, Aiha (2021): *The Constant Boss, Work Under Digital Surveillance*. Data & Society, en https://datasociety.net/wp-content/uploads/2021/05/The_Constant_Boss.pdf.

O'NEIL, Cathy (2017): *Armas de destrucción matemática*. Capitán Swing, Madrid.

PASQUALE, Frank (2019): "A Rule of Persons, Not Machines: The Limits of Legal Automation", The George Washington Law Review, vol. 87, núm. 1, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3135549.

PASQUALE, Frank (2020): New laws of robotics: defending human expertise in the age of AI. The Belknap Press.

PROTASIEWICZ, Jarosław, PEDRYCZ, Witold, KOZŁOWSKI, Marek, DADAS, Sławomir, STANISŁAWEK, Tomasz, KOPACZ, Agata, y GAŁĘŻEWSKA, Małgorzata (2016): "A recommender system of reviewers and experts in reviewing problems". *Knowledge-Based Systems*, vol. 106, pp. 1643178, DOI: 10.1016 j.knosys.2016.05.041.

PUJOL VILA, Oriol (2021): "The concept of 'artificial intelligence'. Opactity and societal impact". 2020. En Pablo García Mexía y Francisco Pérez Bes (eds.): *Artificial Intelligence and the law*. Wolters Kluwer.

RIVAS VALLEJO, Pilar (2020): La aplicación de la inteligencia artificial al trabajo y su impacto discirminatorio. Thomson Reuters Aranzadi, Cizur Menor.

RIVAS VALLEJO, Pilar, dir. (2022): Discriminación algorítmica en el ámbito laboral: perspectiva de género e intervención. Thomson Reuters Aranzadi, Cizur Menor.

ROIG, Antoni (2020): *Las garantías frente a las decisiones automatizadas: del Reglamento General de Protección de Datos a la gobernanza algorítmica*. Bosch, Barcelona.

ROSENBLAT, Álex (2018): Uberland Cómo los algoritmos están reescribiendo las reglas de trabajo. University of California Press.

RUIZ-GALLARDÓN, Isabel (2017): "La equidad: una justicia más justa". *Foro, Nueva época*, vol. 20, núm. 2, pp. 173-191, http://dx.doi.org/10.5209/FORO.59013.

SCANTAMBURLO, Teresa (2021): "Non-empirical problems in fair machine learning". *Ethics and Information Technology*, núm. 23, pp.703–712, https://doi.org/10.1007/s10676-021-09608-9.

SCHIEK, Dagmar y LAWSON, Anna (dirs.) (2016): European Union Non-Discrimination Law and Intersectionality: Investigating the triangle of racial, gender and disability discrimination, Londres-Nueva York: Routledge.

SERRA CRISTÓBAL, Rosario (coord.) (2013): La discriminación múltiple en los ordenamientos jurídicos español y europeo, Valencia: Tirant lo Blanch.

SERRA CRISTÓBAL, Rosario (2020): "El reconocimiento de la discriminación múltiple por los tribunales". *Teoría y derecho*, núm. 27, pp. 140-161. DOI: https://doi.org/10.36151/td.2020.008.

SLAVIN, Kevin (2011): "Cómo los algoritmos configuran nuestro mundo", TED talks, 11/7/2011, https://www.ted.com/talks/kevin_slavin_how_algorithms_shape_our_world?language=es.

SMITH-STROTHER, Lisa (2016): "The role of social advocacy in diversity & inclusion recruiting", Glassdoor Summit 2016, https://youtu.be/IdsqQMV4V_0).

SPIEGELHALTER, David, y HARFORD, Tim (2014): "Big data: are we making a big mistake?" *The Financial Times*, 28/3/2014, en https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0.

STOICA, Ana-Andreea, RIEDERER, Christopher, y CHAINTREAU, Augustin (2018): "Algorithmic glass ceiling in social networks: the effects of social recommendations on network diversity". *Proceedings of the Web Conference* 2018, Lyon. ACM, Nueva York, pp. 923–932, https://doi.org/10.1145/3178876.3186140.

TOMEI, Manuela (2003): "Análisis de los conceptos de discriminación y de igualdad en el trabajo". *Revista Internacional del Trabajo*, vol. 122, núm. 4.

VANTIN, Serena (2021): "Inteligencia artificial y derecho antidiscriminatorio", en LLANO ALONSO, F. y GARRIDO MARTÍN, J. (eds.): *Inteligencia artificial y derecho. El jurista ante los retos de la era digital*. Thomson Reuters Aranzadi, Cizur Menor.

XENIDIS, Raphaële (2021): "Tuning EU equality law to algorithmic discrimination: three pathways to resilience", *Maastricht Journal of European and Comparative Law* 2020, vol. 27, núm. 6, 4/1/2021, en https://doi.org/10.1177/1023263X20982173.

XENIDIS, Raphaële y SENDEN, Linda (2020): "EU non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination". U. Bernitz et al (eds): *General principles of EU law and the eu digital order*. Kluwer Law Int., pp. 151-182.

WACHTER, Sandra (2020): "Affinity profiling and discrimination by association in online behavioural advertising". *Berkeley Technology Law Journal*, núm. 35, https://btlj.org/data/articles2020/35_2/01-Wachter_WEB_03-25-21.pdf.

ZUIDERVEEN BORGESIUS, Frederik (2018): *Discrimination, artificial intelligence, and algorithmic decision-making*. Council of Europe, Directorate General of Democracy.

NOTA: todos los enlaces digitales fueron verificados en el mes de enero de 2022.