ESTUDIOS PENALES Y CRIMINOLÓGICOS

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

Artículos Doctrinales

JUSTICIA ALGORÍTMICA: UN ENFOQUE SOCIOTÉCNICO

ALGORITHMIC JUSTICE: A SOCIOTECHNICAL APPROACH

Andrea Bravo Bolado[1,a]

[1] Contratada predoctoral FPU. Departamento de Derecho Público y Filosofía Jurídica. Facultad de Derecho. C/ Kelsen, 1, 28049, Universidad Autónoma de Madrid, C/ Kelsen, 1, 28049, España

[a] andrea.bravo@uam.es

Resumen

Los instrumentos de predicción del riesgo de reincidencia basados en IA generan numerosos retos para los pilares del Derecho penal. Es urgente una introspección en una disciplina incipiente, el "*algorithmic fairness*", cuyo objetivo es construir herramientas éticas, adaptadas al concepto de "justicia". Se pretende arrojar claridad metodológica en un campo donde convergen disciplinas dispares (ciencia de datos, matemáticas y derecho), intentando dar respuesta a los siguientes interrogantes: ¿es posible trasladar al lenguaje matemático conceptos como "equidad" o "no discriminación"?; ¿hay varios conceptos de lo "justo"? ¿son compatibles? ¿qué resultados arrojan?; ¿es posible tender un puente entre ambos lenguajes que brinde resultados objetivamente más justos?; ¿de qué manera el desarrollo de derechos como la igualdad y la no discriminación deben afectar en la programación? Dicho examen nos permitirá dotar de la necesaria protección a colectivos que sufren el riesgo de ser cada vez más marginalizados por el sistema.

Palabras clave: algoritmos; equidad; justicia predictiva; no discriminación.

Abstract

AI-based tools for predicting the risk of recidivism raise numerous challenges for the grounds of criminal law. There is an urgent need for introspection in an emerging discipline, "algorithmic fairness", which aims to build ethical tools, adapted to the concept of "justice". The aim is to provide methodological clarity in a field where disparate disciplines (data science, mathematics and law) converge, trying to answer the following questions: is it possible to translate concepts such as "fairness" or "non-discrimination" into mathematical language?; are there various concepts of "fairness"?; are they compatible?; what results do they yield?; is it possible to bridge the gap between the two languages to provide objectively fairer results?; how should the development of rights such as equality and non-discrimination affect programming? Such an examination will allow us to provide the necessary protection to groups that are at risk of being increasingly marginalised by the system.

Keywords: algorithms; equity; predictive justice; non-discrimination.

1. INTRODUCCIÓN

La expresión que conforma el título del presente artículo pretende poner sobre aviso al lector acerca del problema y los enfoques que se pretenden aportar en un campo que, a primera vista, parece siempre vasto, complejo e inabarcable. Trataré a lo largo de estas páginas de desentrañar los problemas que la unión de dos disciplinas complejas presenta; y lo haré centrándome en un campo muy estrecho.

En primer lugar, hablando de "justicia algorítmica". El término justicia se utiliza, aquí, en un doble sentido. El primero de ellos; "justicia" en su sentido más institucionalizado, en referencia a la Administración de Justicia que, si bien es cierto que con diferente intensidad en función de la ubicación geográfica, asiste a una transformación tecnológica que parece inevitable. Así, y sin ánimo de exhaustividad, podemos hablar de diversas herramientas que comienzan a revestir nuestro sistema de justicia penal de un aura cientificista, cuando, de pronto, es un *software* el que realiza no sólo tareas de investigación (también conocidas como policía predictiva[1]), sino que también es capaz de detectar la probabilidad de que un denunciante testifique falsamente[2] , o, por centrar el objeto en las concretas herramientas que aquí se van a estudiar, ofrece respuestas dicotómicas sobre la frecuente pregunta: "¿es este individuo propenso a realizar una determinada acción en el futuro[3] ?" Aterrizamos, así, en las herramientas o modelos que guiarán el problema de "justicia" (en sentido ético/filosófico) que aquí se desarrollará.

Las llamadas herramientas de predicción del riesgo (*Risk Asessment Instruments* o *Risk Assessment Tools* [4]), se nutren de grandes cantidades de datos históricos que son

[1] Así, en el campo policial destaca el desarrollo de las llamadas técnicas de *predictive policing*. Véase.: BABUTA, A., & OSWALD, M.: "Data analytics and algorithms in policing in England and Wales: Towards a new policy framework", en *RUSI Occasional Paper*, 2020.

[2] Se trata del software Veripol; QUIJANO-SÁNCHEZ, L., LIBERATORE, F., CAMACHO-COLLADOS, J., & CAMACHO-COLLADOS, M.: "Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police", en *Knowledge-Based Systems*, 149, 2018, pp. 155-168.

[3] Para una visión del desarrollo e implantación de estas herramientas y técnicas en el ámbito penitenciario español, con repaso del recorrido histórico anglosajón, véase.: RIVERA BEIRAS, I.: "Actuarialismo penitenciario. Su recepción en España", en *Revista Crítica Penal y Poder*, nº 9, 2015, pp.102-144. Asimismo, desde una visión más científica: ANDRÉS PUEYO, A., & ECHEBURÚA ODRIOZOLA, E.: "Valoración del riesgo de violencia: instrumentos disponibles e indicaciones de aplicación", en *Psicothema*, Vol. 22, nº 3, 2010, pp. 403-409. Para un acercamiento al movimiento que dio lugar al auge del actuarialismo y sus cambios de paradigma; BRANDARIZ GARCÍA, J. Á.: *El modelo gerencial-actuarial de penalidad: Eficiencia, riesgo y sistema penal*, Madrid, 2016.

[4] Para una revisión en mayor profundidad (sin ánimo de agotar la extensísima bibliografía al respecto), vid.: STEVENSON, M.: "Assessing risk assessment in action", en *Minn. L. Rev.*, 103, 2018, p. 303, MONAHAN, J.: "A jurisprudence of risk assessment: Forecasting harm among prisoners, predators, and patients", en *Virginia Law Review*, 2006, pp. 391-435; y COLLINS E.: "Punishing Risk", en *Georgetown Law Journal*, 107(1), 2018, pp. 57-108. En todo caso, debe aclararse que, aunque aquí se ha optado por el uso de la palabra riesgo, como traducción literal de la palabra "risk", de aquí en adelante se utilizará de manera indistinta los conceptos predicción del riesgo y predicción de la peligrosidad a la hora de hablar de estas herramientas. Debe apuntarse, sin embargo, que entre ambos hay una diferenciación marcada, pues, la valoración de la peligrosidad nace asociada al juicio clínico, incluso a factores de corte psicológico e individual, mientras que la valoración del riesgo es una apuesta de quienes consideran que el término peligrosidad goza de imprecisión y subjetivismo, optando por un enfoque basado en métodos actuariales estructurados, pretendidamente científicos y objetivos. Sin embargo, y pese a ser conocedora de la interesante discusión que en cuanto al binomio "peligrosidad/riesgo" se suscita, la autora, consciente de que autores como

seleccionados, filtrados y procesados, para ser introducidos en modelos (cuya complejidad va evolucionando con la misma evolución de la IA y del *machine learning* [5]), y son entrenados con el objetivo de clasificar, atendiendo a ciertos parámetros de optimización, a los individuos, en función de su probabilidad de volver a cometer un hecho (ya sea de volver a cometer un delito, de escapar de la acción de la justicia, o de dañar bienes concretos de la víctima). Lejos queda ya el paradigma que dio lugar a las primeras predicciones algorítmicas en el campo de la justicia penal, pues son muy numerosas las herramientas que han ido desarrollándose e implantándose en los diversos sistemas penales de nuestro entorno, desde el ámbito anglosajón (donde mayor desarrollo y debate entrañan), hasta el entorno europeo, pasando, por supuesto, por las gigantes potencias asiáticas que, con China a la cabeza, se han ido sumando a la tendencia que todo lo arrasa; la confianza en el dato como fuente de información omnipotente y omnipresente.

Con este desarrollo imparable, parece que la legislación se ha visto sorprendida ante una realidad en expansión, con una potencialidad lesiva que se extiende a numerosos campos, pero que no ha acabado de ser abordada de una manera uniforme por el derecho (en ocasiones, ni siquiera regulada en absoluto), y que despierta preocupaciones muy razonables entre una comunidad de científicos sociales cada vez más concernida por los desequilibrios que la tecnología pueda introducir en una sociedad que de por sí, tiene la desigualdad y la injustica ínsita en su ADN.

Y esto nos lleva a la segunda acepción del término "justicia" que aquí se pretende exponer. Justicia como valor supremo, principio ético y programático que debe guiar no solo los ordenamientos jurídico en un sentido teórico, sino también tener un reflejo práctico en los resultados y consecuencias de todo un sistema en el que los derechos fundamentales más esenciales están en juego.

Haciendo un breve estudio de los instrumentos normativos[6] que, al albor de las nuevas tecnologías asociadas a la IA, se han ido aprobando en nuestro entorno, con especial atención a los instrumentos europeos que nos aportan una visión común, es fácil divisar que una de las

---

Rigakos o Martínez Garay lo ponen de manifiesto, considera que la peligrosidad sigue teniendo un papel protagónico en la legislación penal, incluso allí donde se aplican las modernas herramientas de medición del riesgo, por lo que opta por utilizar ambos términos, al entender que la pretendida objetividad de quienes propugnan la sola presencia de la "evaluación del riesgo" no puede darse por sentada y válida sin mayores pretensiones. Vid., en lo relativo a la discusión terminológica: ANDRÉS PUEYO, A., REDONDO ILLESCAS, S., "Predicción de la violencia: entre la peligrosidad y la valoración del riesgo de violencia", en P*apeles del Psicólogo*, vol. 28, núm.3, septiembre-diciembre 2007, pp. 157-173; MARTÍNEZ GARAY, L., MONTES SUAY, F.: "El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias" en *InDret*, 2/2018; ANDRÉS-PUEYO, A., "Peligrosidad criminal: análisis crítico de un concepto polisémico, en MAROTO CALATAYUD y DEMETRIO CRESPO (Coords.), "Neurociencias y derecho penal: nuevas perspectivas en el ámbito de la culpabilidad y tratamiento jurídico-penal de la peligrosidad", ed. Edisofer, 2013; RIGAKOS, G.S., "Risk society and actuarial criminology: prospect for critical discourse", en *Canadian Journal of Criminology*, 41 (2), 1999, pp. 137-150.

[5] Lo cierto es que las definiciones formales o jurídicas de estas novedosas técnicas, precisamente por su constante evolución y difícil aprehensión teórica, son una de las carencias de los instrumentos normativos actuales. Aun así, podemos encontrar algunas definiciones muy genéricas en instrumentos normativos de *soft law* a nivel europeo. Véase la ofrecida por el Grupo Independientes de Expertos de Alto Nivel sobre Inteligencia Artificial, creado por la Comisión Europea en junio de 2018, en su documento con las Directrices éticas para una IA fiable, en el que se definen los sistemas de IA como "*sistemas de software (y posiblemente también de hardware) diseñados por humanos que, dado un objetivo complejo, actúan en una dimensión física o digital percibiendo su entorno a través de la obtención de datos, interpretando los datos estructurados o no estructurados recabados, razonando sobre el conocimiento o procesando la información derivada de dichos datos y decidiendo las mejores acciones a tomar para conseguir el objetivo*".

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

preocupaciones más patentes es la relativa a la "justicia" de las herramientas de Inteligencia Artificial[7] . Así, se considera esencial la búsqueda de algoritmos "justos"[8] , cuyos resultados o procesos no supongan una violación de los principios de igualdad y no discriminación, que no estén sesgados, y que traten de evitar repetir aquellos errores del pasado reciente que han mostrado cómo los colectivos tradicionalmente sometidos a discriminaciones intolerables pueden verse doblemente dañados cuando dichas técnicas entran en juego.

Resulta ilustrativo el análisis que sobre la cuestión de la justicia aquí planteada realiza el grupo independiente de expertos de alto nivel sobre IA en sus Directrices éticas para una IA fiable. Partiendo del objetivo de la UE de obtener una IA fiable, se desarrolla la necesidad de una IA ética, con la que se respeten los derechos fundamentales reconocidos por la Unión y de obligado cumplimiento para todos sus componentes.

Se habla a lo largo del documento de igualdad, no discriminación y solidaridad. En el contexto de la IA, la igualdad implica que el funcionamiento de este tipo de sistemas no debe generar resultados injustamente sesgados y esto requiere, a su vez, un adecuado respeto de las personas y grupos potencialmente vulnerables, como los trabajadores, las mujeres, las personas con discapacidad, las minorías étnicas, los niños, los consumidores u otras personas en riesgo de exclusión.

Los principios esenciales de los que parte el documento (respeto a la autonomía humana, prevención del daño, equidad y explicabilidad) se asocian con derechos protegidos en la Carta de Derechos fundamentales de la UE. Así, el respeto de la autonomía humana está estrechamente relacionado con el derecho a la dignidad y la libertad humanas (recogido en los artículos 1 y 6 de la Carta), la prevención del daño a la protección de la integridad física o mental (reflejada en el artículo 3), la equidad a los derechos a la no discriminación, la solidaridad y la justicia (recogidos en el artículo 21 y siguientes) y la explicabilidad y responsabilidad están relacionadas, a su vez, con los derechos referentes a la justicia (reflejados en el artículo 47).

De esta forma, podemos apreciar cuál es el camino que marcan ya, si bien de una manera genérica, los instrumentos disponibles, y tal será el método de aproximación que se siga a lo largo de estas páginas, aunando (en la manera también apuntada por Aizenberg y Van den

---

[6] Por nombrar algunos de los más significativos a nivel europeo: Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno adoptado por el CEPEJ durante su 31ª Reunión plenaria, adoptada en Estrasburgo los días 3 y 4 de diciembre de 2018. Directrices éticas para una IA fiable Grupo de expertos de alto nivel sobre inteligencia artificial, de 8 de abril de 2019. 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; el Libro Blanco sobre la Inteligencia artificial aprobado por la Comisión en febrero de 2020, y 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, aprobada en abril de 2021. Para una visión más completa de las directrices éticas que, a nivel internacional, se han aprobado en los últimos años, véase: VALLS PRIETO, J.: *Inteligencia artificial, Derechos Humanos y bienes jurídicos*, Thomson Reuters, Aranzadi, 2021, pp. 71 y ss.

[7] Para una revisión interactiva de los diferentes principios normativos a nivel internacional, se puede consultar la herramienta "*AI Principles Map*", desarrollada por el AI Ethics Lab, que muestra un mapa interactivo donde se recogen las tendencias en materia de ética regulatoria en los diferentes estados del mundo, diferenciando no solo en función de los principios protegidos, sino de las entidades de los que provienen. Enlace: https://aiethicslab.com/big-picture/ (consultado por última vez el 9/11/2022).

[8] Se hace necesario, aquí, realizar una precisión terminológica, pues al traducir los textos y conceptos de la literatura, principalmente en lengua inglesa, se ha optado por usar indistintamente las palabras "equidad" y "justicia" como equivalentes para el término inglés "*fairness*". Sin embargo, se debe llamar la atención sobre el hecho de que no todos los autores de lengua hispana acuñan estos mismos términos.

Hoven[9]) los principios éticos y jurídicos que sirven de guía a nuestras sociedades, con los problemas técnicos que están ya sobre la mesa.

En definitiva, y cerrando esta introducción con la alusión a la segunda frase del encabezamiento, se hace necesario aquí un enfoque "sociotécnico[10] " para abordar la problemática de la búsqueda de justicia dentro de la Administración de Justicia (concretamente, penal). Me tomaré la licencia de hacer una bipartición más, en aras a una mayor claridad. Se necesita un enfoque capaz de aunar dos mundos, tradicionalmente concebidos como cajones estancos, que, de una vez por todas, deben convertirse en vasos comunicantes. Lo social y lo tecnológico deben ir de la mano en la búsqueda del ideal de justicia; científicos de la computación y juristas deben intercambiar un diálogo donde ambas voces se escuchen, donde ambas visiones aporten, pues solo así podremos reflexionar coherentemente sobre aquello que es importante, aquello que es posible, y aquello que debe entrar a debatirse.

Los interrogantes a los que se tratará de dar respuesta son, por tanto, del siguiente tipo: ¿es posible trasladar al lenguaje matemático conceptos como "equidad" o "no discriminación" ?; ¿hay varios conceptos de lo "justo"? ¿son compatibles? ¿qué resultados arrojan?; ¿es posible tender un puente entre ambos lenguajes que brinde resultados objetivamente más justos?; ¿de qué manera el desarrollo de derechos como la igualdad y la no discriminación deben afectar en la programación?

El propósito de este trabajo es, pues, tratar de aunar las voces de estos tradicionales contrincantes, para poner al servicio de la comunidad una reflexión que nos incumbe a todos; ¿cuál es el camino para conseguir una tecnología justa en el concreto contexto del sistema penal? Veamos, pues, de dónde partimos.

---

[9] AIZENBERG, E., & VAN DEN HOVEN, J.: "Designing for human rights in AI", en *Big Data & Society*, July–December 2020, pp. 1-14.

[10] No puede entrarse aquí a reflexionar en profundidad sobre la totalidad de la influencia que los estudios sociales sobre ciencia y tecnología han aportado como campo de investigación. Baste dejar apuntada su valiosa contribución en la construcción de un concepto crítico e interdisciplinar, enfocado en los problemas éticos que buscan entender cómo ciencia y técnica afectan a la sociedad en diversidad de campos, considerando los aspectos éticos y morales de su desarrollo. Autores como Habermas o Foucault, así como importantes pensadores actuales tales como Latour, Haraway, Feenberg o Byun Chul Han son referentes en este campo. Sin embargo, las referencias que, de forma ya muy centrada en el objeto de discusión del presente artículo, se centran en la concreta aplicación de la manida problemática al campo de la equidad de los más sofisticados algoritmos de IA, vienen de la mano de autores que, desde la ciencia o la filosofía, aplican esta corriente "sociotécnica" a problemas relacionados directamente con lo que aquí se trata; pues ponen el énfasis en lo particular de esta tecnología en concreto en una sociedad también concreta. No se trata de analizar la técnica y la sociedad en un sentido genérico, reflexión que desborda con mucho las pretensiones de este trabajo. Para mayor profundidad, vid.: LATOUR, B., "Reensamblar lo social: una introducción a la teoría del actor-red", Buenos Aires, Manantial, 2008; FEENBERG, A., *Transforming technology: A critical theory revisited* / (2 ed.), Nueva York, Oxford University Press, 2002; HARAWAY, D., *Manifiesto para Cyborgs: [ciencia, tecnología y feminismo socialista a finales del siglo XX]*, Letra Sudaca Ediciones, Buenos Aires, 2018; HAN, *En el enjambre* (1a edición.). Herder, Barcelona, 2014 y, más específicamente centrado en el problema que nos ocupa: SELBST, A. D., BOYD, D., FRIEDLER, S. A., VENKATASUBRAMANIAN, S., & VERTESI, J.: Fairness and abstraction in sociotechnical systems", en *Proceedings of the conference on fairness, accountability, and transparency*, January 2019, pp. 59-68. En este último, se parte de la importancia de entender los concretos sistemas que son objeto del presente artículo como "combinación de componentes sociales y técnicos". El propio grupo de Expertos de Alto Nivel acoge dicho concepto en este concreto campo al afirmar que: "*Las directrices, que van dirigidas a todas las partes interesadas, buscan ofrecer algo más que una simple lista de principios éticos; para ello, proporcionan orientación sobre cómo poner en práctica esos principios en los sistemas sociotécnicos*".

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

2. LA PARTE TECNOLÓGICA: LA INCIPIENTE DISCIPLINA DEL *ALGORITHMIC FAIRNESS*

Se comenzará exponiendo el necesario punto de vista técnico, que debe servir como marco esencial a los juristas, al que debemos prestar especial atención y que viene a dar respuesta a algunas de las preguntas que los científicos sociales tradicionalmente se plantean.

El *algorithmic fairness* [11] nace como especialidad cuando la comunidad científica, con el desarrollo de herramientas de Inteligencia Artificial y *machine learning*, comienza a ser consciente de los riesgos asociados a un uso no precavido de los modelos.

Así, en palabras de Zliobaite, "*el machine learning y la minería de datos con conciencia de la discriminación (discrimination-aware) es una disciplina emergente que estudia el potencial discriminador de los algoritmos, y la transparencia y responsabilidad del uso de algoritmos para la toma de decisiones. La disciplina se apoya en el diseño de algoritmos mediante machine learning y técnicas de minería de datos, la ciencia estadística, así como en la sociología y la ley para definir la justicia. El objetivo hasta ahora ha sido desarrollar técnicas algorítmicas que pudieran obedecer las regulaciones de justicia prescritas por la ley. El foco ha estado puesto en dos retos principales: cómo formular restricciones de justicia matemáticamente y cómo hacer que los modelos las obedezcan"* [12] .

Los científicos de datos se han encargado de desentrañar el complejo proceso de procesamiento de datos, diseccionándolo en sucesivos pasos, y señalando de forma detallada cuáles son las principales fuentes de peligro, así como advirtiendo sobre dónde se debe poner el foco si se quiere mitigar la proliferación de errores. Se utilizará, a continuación, para ofrecer una explicación uniforme, la lógica común de todos los métodos de *machine learning* supervisado, dejando de lado todas las técnicas de aprendizaje autónomo o no supervisado que, si bien no paran de proliferar en su desarrollo, encuentran aplicación, al menos por el momento, en campos distintos.

Así, ofreciendo una breve caracterización; los modelos de predicción, entre los que se encuentran las herramientas de predicción de la peligrosidad, son modelos matemáticos que predicen un resultado a partir de las características de un objeto. El objetivo es alcanzar la mejor precisión posible en los datos nuevos, aquellos sobre los que el modelo no se ha entrenado y no conoce.

En esencia, el proceso de creación de un modelo de predicción se basa en tres grandes estadios: la recolección de datos, el procesamiento de estos para entrenar al modelo, y la evaluación de los resultados obtenidos[13] . A lo largo de estas tres fases, diversas complicaciones pueden dar lugar a diferentes resultados problemáticos. No se tratará, a continuación, de dar una calificación jurídica a las posibles desviaciones del modelo, sino simplemente de exponer qué puede ocurrir en cada uno de estos estadios.

---

[11] Se utilizará, a lo largo del texto, su denominación en inglés para hacer alusión a la disciplina apegada a la ciencia de datos, y no dar lugar a confusión con otros términos similares que podrían derivarse de la traducción de esta expresión, aunque las denominaciones de esta disciplina son variadas. Así, *AI fairness, fairness in AI o ML fairness* son otras de las más utilizadas entre los científicos de datos, todas ellas conteniendo la palabra "*fairness*" como central.

[12] ZLIOBAITE, I.: "Fairness-aware machine learning: a perspective" en *arXiv:1708.00754*, 2017, p. 3[traducción de la autora].

[13] CALDERS, T., & ŽLIOBAITĖ, I.: "Why unbiased computational processes can lead to discriminative decision procedures", en CUSTERS B., CALDERS, T., SCHERMER, B, & ZARSKY, T. (Eds.), *Discrimination and Privacy in the Information Society*, Springer, Berlin, Heidelberg, 2013, pp. 45-46.

Antes de entrar a desentrañar el proceso técnico, se hace necesario aclarar el significado de un término que aparece continuamente en la literatura científica (y que se ha acabado colando, inevitablemente, en la jurídica), cuando se habla de problemas de justicia; el sesgo. El sesgo es una tendencia o inclinación dotada de un significado moral, pues implica la discriminación sistemática e injusta en contra de ciertos individuos o grupos de individuos en favor de otros[14] .

La detección y mitigación de sesgos es, pues, el objetivo primordial y piedra angular del *algorithmic fairness*. Veamos, ahora sí, cómo se desarrolla este proceso, que podemos diseminar, siguiendo el esquema propuesto por Suresh y Gutag[15] , en los siguientes pasos:

a) El proceso empieza con la recolección de datos. Este proceso implica definir una población (*target population*) y tomar una muestra de la misma, así como identificar y medir las características y marcadores (*labels*). Este *dataset* se dividirá entre los datos de entrenamiento y los datos de testeo.

b) El modelo es definido y optimizado con los datos de entrenamiento.

c) El conjunto de datos de testeo se usa para la evaluación del modelo, y el modelo final es integrado en el contexto del mundo real de que se trate.

A continuación, se examinarán los posibles momentos en que los errores en las diferentes fases pueden dar lugar a sesgos injustos. Aunque, y tal y como explican los autores, existen diversas formas de categorizar y sistematizar los sesgos (por ejemplo, en función de si interactúan con los datos, con el algoritmo o con el usuario), se adoptará aquí el enfoque que se centra en los diferentes momentos del ciclo de vida del algoritmo, por considerarse uno de los métodos más ilustrativos para las personas que comienzan a acercarse a la dinámica de funcionamiento de los sistemas de *machine learning*. No en todos los escritos científicos se utilizan exactamente las mismas denominaciones para estos sesgos (por ejemplo, Nissembaum[16] utiliza el término sesgo "preexistente" para referirse al sesgo "histórico" según la denominación de Suresh y Gutag). Pero pese a la divergencia terminología, es posible y deseable realizar una aproximación metodológica a estas clasificaciones.

2.1 Recolección de datos y su procesamiento

2.1.1 Sesgo histórico o preexistente

Podríamos afirmar que este error ocurre incluso en un estadio previo a la misma recolección o creación del modelo, pues se produce cuando los datos, aunque sean perfectamente recogidos y medidos, suponen una representación del mundo tal y como es, de forma que dichos datos reflejan también los perjuicios y resultados injustos. Los autores arriba citados ponen el ejemplo de las *corpora* (bases de datos masivas) de palabras, llamadas *word-embedding*, usadas para aplicaciones que tienen que ver con el procesamiento del lenguaje natural, y que reflejan, tal y como recogen del mundo real, un lenguaje

[14] FRIEDMAN, B. & NISSEMBAUM, H.: "Bias in computer systems," en *ACM Transactions on Information Systems*, July, vol.14, nº 3, 1996, p. 332.

[15] SURESH, H., & GUTTAG, J.: "A framework for understanding sources of harm throughout the machine learning life cycle", en *Equity and access in algorithms, mechanisms, and optimization*, (EAAMO '21), October. 5-9,2021, p. 3.

[16] *Op. cit*. n. 14.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

inevitablemente machista. Así, si existen subgrupos tradicionalmente infrarrepresentados en según qué datos, como consecuencia de injusticias que hunden sus raíces en instituciones, prácticas y actitudes previas al modelo, esta realidad habrá de ser tenida en cuenta previamente a la creación.

2.1.2 Error de medida (measurement error)[17]

Podemos encontrarnos con dos errores de medida típicos, que dan lugar a dos sesgos diferentes; su denominador común viene dado por una discrepancia entre la realidad a observar y la representación que de ella se hace en los datos. Así, como consecuencia de esta disparidad entre realidad y datos, observamos dos posibles errores:

2.1.2.1 Sesgo en los atributos (*feature bias*); hay una discrepancia entre los datos que se pretenden recoger como indicadores o atributos y los que realmente están disponibles o es posible medir. Por ejemplo, que, de cara a diseñar una herramienta de predicción del riesgo de reincidencia, uno de los atributos o características recogidas sea el número de veces que una persona ha cometido un delito implica un problema de este tipo, pues se da una discrepancia entre la realidad y la medición. ¿Cómo se puede medir de manera fidedigna si alguien ha cometido un delito? La única manera será hacerlo mediante la recogida de indicadores (*features*) imperfectos, como es el caso del número de veces que una persona es detenida o procesada por un delito. Se desconoce, así, que no hay una identificación total entre ambas realidades, pues muchos delitos tienen una cifra negra que, y de forma muy diferenciada según la tipología delictiva, hace que no haya una correspondencia entre el delito y su descubrimiento[18] .

2.1.2.2 Sesgo en la etiqueta (*label bias*): en este caso, la discrepancia se produce entre los datos que se recogen y los resultados que se utilizan como medidor de corrección. Es decir, una vez que se etiqueta un dato bruto y esta etiqueta (*label*) se utiliza como indicador de un rasgo, puede ocurrir, una vez más, que dicho rasgo no capte la realidad en su totalidad. No es posible aprehender de forma perfecta ese rasgo, y se acude a un indicador próximo (denominado en el lenguaje técnico *proxy*) que correlaciona con él pero que no es equivalente. El caso del arresto es, una vez más, problemático, esta vez utilizado como indicador del resultado; es decir, si se quiere medir si el resultado del modelo es correcto se podrá decidir que el indicador del riesgo es que una persona haya vuelto a cometer un delito, pero, una vez más, como este indicador o etiqueta

[17] VEALE, M., & BINNS, R.: "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data", en *Big Data & Society*, July–December, 2017, pp. 1-17; GOEL, S., SHROFF, R., SKEEM, J., & SLOBOGIN, C.: "The accuracy, equity, and jurisprudence of criminal risk assessment", en VOGEL, R. (Ed). *Research handbook on big data law. Intellectual Property Forum: Journal of the Intellectual and Industrial Property Society of Australia and New Zealand*, ed. Edward Elgar Publishing, 2021, pp. 9-28.

[18] La solución que algunos autores proponen pasa por elegir, en la medida de lo posible, o bien atributos que admitan una medición más objetiva o bien atributos que presenten una mayor concordancia con la realidad. Por ejemplo, será más fácil que no haya tanta cifra negra en los delitos violentos como un asesinato o un homicidio, que en delitos de menor envergadura o aquellos cometidos en contexto muy propensos a la no revelación, como es el caso de los abusos a menores o la ciberdelincuencia. Véase: CALDERS, T., & ŽLIOBAITĖ, I., *op. cit*. n. 13.

> es difícil de recoger y medir, se utiliza un indicador próximo, el arresto. Pero podemos encontrar otros ejemplos fuera de la predicción de la peligrosidad criminal. Véase la utilización de factores como la evaluación de capacidades en el marco de la selección de personas para puestos de trabajo. Si se quiere medir el éxito laboral como indicador de que una persona tenga ciertas aptitudes necesarias para el puesto en cuestión, dicho indicador es de difícil o imposible aprehensión objetiva. Habrá que acudir, pues, a indicadores próximos (*proxies*) del éxito, por ejemplo, haber sido objeto de ascensos en el pasado. Una vez más, no hay una relación de perfecta concordancia entre aquello que se quiere medir y lo que en realidad se mide.

Otros autores como Veale y Binns[19] se refieren a una realidad parecida bajo la denominación "*feature engineering*", que podríamos traducir como ingeniería de atributos. En este caso, lo que ocurre es que los datos son manipulados, limpiados o transformados por los operadores humanos, introduciendo alteraciones subjetivas que alteran la medición. Por ejemplo, unificar a todas las personas creyentes en diferentes ramas de religión (católicos, protestantes, chiitas y sunitas), dentro de su doctrina general (cristiano/musulmán), puede hacer que se pasen por alto algunas distinciones que son muy relevantes de cara a alcanzar la justicia en determinados contextos.

2.1.3 Sesgo en la muestra (sample bias) o sesgo de representación (representation bias)

Los datos recogidos como muestra no son suficientemente representativos de la población sobre la que se va a aplicar el modelo. Este es un error sencillo de comprender, y, en principio, que tiene una solución más fácil que la de los anteriormente expuestos. Se trata de que la muestra de datos sea lo más parecida a la población sobre la que se va a aplicar el modelo en el futuro, pues solo así recogerá las particularidades necesarias para una optimización en las predicciones. Aun así, observamos cómo ha sido uno de los errores más típicamente detectados en el proceso de implantación de modelos. Ocurre, por supuesto, también en el campo de la predicción de la reincidencia, pues en ocasiones herramientas creadas para su implantación en contextos geográficos con características muy peculiares se han trasladado a regiones que nada tienen que ver, y cuya realidad delincuencial hace que los factores de predicción no funcionen en iguales condiciones[20] . Sucede también este error en contextos médicos. Por ejemplo, exponen Lin y Chen que, en el ámbito médico estadounidense, muchas bases de datos médicas son "europeo céntricas" y que carecen de una geodiversidad necesaria, pues están sesgadas a favor de las poblaciones más privilegiadas, de forma que, en el caso de la detección de cáncer de piel, los *dataset* utilizados son mayoritariamente de gente con piel blanca, por lo que, para las personas negras, infrarrepresentadas en los modelos de entrenamiento, el resultado de su diagnóstico resulta ser mucho menos preciso[21] .

[19] VEALE, M., & BINNS, R., op cit. n.17, p. 2.

[20] Es el caso del sistema ORAS (*Ohio Risk Assessment System*), que fue creado sobre una muestra de población especialmente pequeña correspondiente al Estado de Ohio, pero que acabó aplicándose a nivel nacional.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

2.2 Entrenamiento del modelo

Es en este momento donde, una vez recolectados y, en su caso, procesados los datos (ya sea balanceando la muestra para que resulte adecuada a la población sobre la que se usará el modelo, o modificando algunos indicadores para que sean más objetivos), debe ponerse en marcha el proceso de computación propiamente dicho. No se entrará aquí a explicitar el funcionamiento de las diferentes técnicas, pues cada una de ellas tiene una complejidad propia, que, por otra parte, no aporta demasiado al debate que aquí nos ocupa (sí lo hace, sin embargo, en otros debates colindantes, como el relativo a la explicabilidad de las herramientas, otro de los grandes pilares fundamentales en el desarrollo de la IA). En lo que interesa, cabe afirmar que el denominador común de todos estos modelos predictivos es el de realizar una tarea clasificatoria. Es decir, el objetivo inherente a los modelos algorítmicos, utilicen la técnica que utilicen, es discriminar. Se fijan ciertos parámetros de corrección, cuyo objetivo se basa en minimizar el error (en cualquiera de sus posibles variantes o parámetros), dando lugar a un resultado (*output*), que determina cuál es la división óptima dados los datos y los indicadores disponibles.

Así, y trasladando esta explicación abstracta al concreto campo de la predicción de la reincidencia, la tarea asignada al modelo es discriminar cuáles son los indicadores que pueden marcar mejor la línea divisoria entre aquellas personas propensas a delinquir o aquellas que no lo son.

La manera en que los modelos llegan a este resultado es la iteración. Una y otra vez los mismos datos son presentados al modelo, que va recalculando el error hasta llegar a un punto óptimo. Esta dinámica tiene un efecto perverso, pues por su propia naturaleza implica que los mismos datos (con sus errores ínsitos), son presentados como correctos, y sobre ellos aprende el modelo, infiriendo características que, en ocasiones, son producto no de hechos objetivos, sino puramente valorativos. Así, si un modelo se nutre de los resultados de una entrevista realizada por una persona que tiene un sesgo machista, propensa a no contratar a mujeres simplemente por el hecho de serlo, e incluso aunque este sesgo sea inconsciente para el entrevistador, el modelo interpretará que aquellas personas que reúnan esa condición (sea explícito o no el factor del sexo como atributo), tendrán que caer del lado de la clasificación negativa. El problema de los *feedback loops* (o bucles de retroalimentación), es, pues, muy relevante cuando los datos históricos de los que se parte presentan estos sesgos. Es este un fenómeno ampliamente reconocido por la comunidad científica: *"Los perfiles empezarán a normalizar la población de la que se extrae la norma, [de tal forma que el modelo] se ajusta a un patrón; el patrón se retroalimenta con las opciones fijadas por el mismo patrón; estas opciones refuerzan al patrón; y el ciclo comienza otra vez"* [22] .

Resulta esencial, pues, detectar y mitigar estos sesgos previamente, pues una vez que llegan a la fase de procesamiento, dada la dinámica reiterativa propia de los modelos, será prácticamente imposible evitar sus efectos perniciosos.

En este momento cobra importancia también un problema de tipo técnico, que Suresh y Gutag denominan "sesgo de aprendizaje" (*learning bias*). Afirman que una de las decisiones más importantes es la de la función (en sentido matemático) que el algoritmo aprende y que

[21] LIN, T.A., & CAMERON CHEN, P. H..: "Artificial Intelligence in a Structurally Unjust Society", disponible en número futuro de *Feminist Philosophy Quarterly*, disponible en la actualidad en: https://philpapers.org/rec/LINAII-2 (consultado por última vez el 9/11/2022), p. 7.

[22] LESSIG, L., *Code: and Other Laws of Cyberspace*, Version 2.0, New York, NY: Basic Books, 2006, p. 220 [traducción de la autora].

tiene que optimizar durante este proceso de entrenamiento reiterativo. Típicamente, estas funciones codifican algún tipo de medida de precisión (por ejemplo, y sin entrar en mayores complejidades, la minimización de algún parámetro de error, como puede ser la entropía cruzada o el error cuadrático medio). Ocurre que, inevitablemente, al elegir las características técnicas del modelo, se está tomando una decisión sobre otros valores en juego, pues priorizar un objetivo (como la precisión global), daña otros (dando lugar a una mayor disparidad). Este tradicional *trade-off* entre precisión y equidad, magníficamente explicado por los científicos Micheal Keans y Aaron Roth, requiere, en primer lugar, ser conscientes de los puntos de equilibrio dentro de los cuales será necesario moverse[23] .

2.3 Evaluación del modelo

Es este uno de los momentos donde mayores tensiones con el principio de justicia o equidad se producen, y donde se han generado, hasta el momento, las mayores batallas y debates entre la comunidad científica y la jurídica.

Puede surgir aquí el denominado "sesgo de evaluación", que surge por el deseo de comparar cuantitativamente los modelos, de forma que es necesario elegir las métricas capaces de evaluar el rendimiento. Se hace necesario, pues, evaluar la capacidad de rendimiento de cada métrica.

Es precisamente aquí donde comienza un debate de muy difícil solución. Para poder afirmar que un modelo es justo o injusto debemos dotarnos de medidores de esa justicia. Los medidores o parámetros propios de la disciplina jurídica serán explorados en el siguiente apartado de este artículo, para centraros, a continuación, en los parámetros técnicos.

Como apuntan Kehl, Guo y Kessler[24] , argumentar que los algoritmos de valoración de riesgo deben ser elaborados de una manera justa es incontrovertido, pero encontrar la definición concreta de justicia es difícil de precisar. La cuestión de si un algoritmo es técnicamente justo depende mucho de sus objetivos, y da lugar a un gran número de consideraciones normativas. Se puede argumentar que un algoritmo es justo siempre que haga predicciones precisas y consistentes. Sin embargo, la comunidad científica no ha alcanzado un consenso sobre la definición exacta de la "justicia" en el contexto estadístico.

Como apunte previo, y aunque este sí es un lenguaje con el que el jurista estará más familiarizado, se debe introducir otra noción esencial: "los atributos protegidos". Los atributos protegidos son aquellas características o factores susceptibles de ser medidos y cuya utilización, por razones diversas (normalmente razones de discriminación históricas), es vista con ojos críticos. Los más típicos son la raza y el sexo, pero la lista va variando según las

[23] KEARNS M., & ROTH, A.: *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*, Oxford University Press, 2020, p. 69 y ss. Explican estos autores que al construir un modelo sería posible enumerar todas las posibles parejas de puntuación (es decir, a tanto error, tanta inequidad), y podríamos llegar a la conclusión de que sólo algunos modelos son válidos de partida; aquellos en los que se puede mejorar la precisión (o la equidad), sin dañar la otra medida. El nombre de este límite es la denominada curva de Pareto, que establece cuáles son las opciones razonables en estos términos de equilibrio. Sin embargo, a partir de este punto, la decisión humana es inevitable, pues es ella la que debe asignar un "peso" relativo a cada valor (un peso al error y un peso a la justicia o equidad); de forma que se elija aquel punto en el que se considere que la relación error/equidad es adecuada (que, por ejemplo, X errores son aceptables en tanto X equidad sea alcanzada). Lo único que pueden decir los científicos a este respecto es si este punto de corte está dentro de la curva de Pareto, pues de lo contrario se tratará de un mal modelo.

[24] KEHL, D., GUO P., KESSLER S.: "Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing", en *Responsive Communities Initiative, Berkman Klein Center for Internet & Society*, Harvard Law School, 2017, p.30.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

sensibilidades y percepciones jurídico/sociales evolucionan. Por tanto, cuando el científico de datos hace referencia a los "atributos protegidos", se está refiriendo, simplemente, a todos aquellos factores especialmente protegidos en la legislación antidiscriminación.

Un gran número de nociones o criterios formales de justicia se han ido descubriendo dentro de la disciplina del *algorithmic fairness* en los últimos años, en el afán de operacionalizar diferentes concepciones de la igualdad a un nivel matemático[25] . Resulta interesante analizar las más importantes, así como sus limitaciones. Son principalmente tres las posibles métricas presentadas por la comunidad científica:

2.3.1 Anti-clasificación

Implica la no utilización de ninguno de los atributos protegidos en el desarrollo y entrenamiento de los modelos, pues se considera que resulta altamente problemático tratar con este tipo de datos (no solo por ser considerados datos "sensibles" según la normativa de protección de datos, sino por sus potenciales resultados lesivos[26]).

Sin embargo, es casi unánime la respuesta de la comunidad científica frente a esta aproximación calificada de "*naïve*"[27] por los expertos; la no utilización de estos atributos protegidos no hará que los mismos no acaben saliendo a la luz. De hecho, tal y como pone de manifiesto Xenidis[28] , puede incluso hacer la detección de la discriminación mucho más difícil al prescindir de las etiquetas que se usan para visualizar y medir el sesgo. Así, Goel y Shroff[29] afirman que en algunos casos será necesario que los algoritmos de evaluación de riesgo consideren explícitamente las características protegidas para alcanzar resultados equitativos. De hecho, estos últimos autores apuestan precisamente por usar estos atributos protegidos para solucionar alguno de los sesgos arriba estudiados.

Carece especialmente de sentido acudir a esta política de prohibición de uso de datos por un fenómeno conocido como "*proxy discrimination*". Hacker[30] explica cómo, en general, la discriminación por *proxy* se refiere a situaciones en que la información precisa sobre un rasgo no está disponible y se sustituye el parámetro deseado con el que es posible observar. Por ejemplo, si un algoritmo que calcula las primas de seguro encuentra que las personas

---

[25] HACKER, P.: "Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law", en *Common Market Law Review*, 55(4), 2018, p. 1175.

[26] En este sentido, VEALE, M., & BINNS, R, *op. cit.* n. 19, p. 5, conscientes de las constricciones que el RGPD implica respecto de estos datos, abogan por un uso adecuado, proponiendo diversos procedimientos que impliquen a terceras partes confiables conocedoras de los mismos, con el objetivo de alcanzar los objetivos de justicia.

[27] Entre otros: KIM, P. T.: "Auditing algorithms for discrimination", en *University of Pennsylvania Law Review Online*, nº 166, 2017, p. 193: "*some of the technical strategies for nondiscrimination require awareness and use of protected characteristics in order to ensure fairness across groups*". Asimismo.: DWORK C., et al.: "Fairness Through Awareness", en Proc. *3rd Innovations Theoretical Computer Science. Conf.*, 2012. Tambien: VEALE, M., & BINNS, R., op. cit. n. 19.

[28] XENIDIS, R.: "Tuning EU equality law to algorithmic discrimination: Three pathways to resilience", en *Maastricht Journal of European and Comparative Law*, *27*(6), 2020, p. 745, nota a pie de página nº 49.

[29] GOEL, S., SHROFF, R., SKEEM, J., & SLOBOGIN, C., op. cit. n. 17, p. 16. Para solucionar el sesgo de etiquetado (*feature bias*), estos autores apuestan por la creación de herramientas separadas para los grupos específicos; por ejemplo, dado que los hombres y las mujeres reinciden en diferente medida, sería aconsejable que su género fuera tomado para realizar modelos separados, de forma que no se produjeran resultados injustos pues, de lo contrario, una puntuación de riesgo que fuera neutra en cuanto al género sobreestimaría de forma sistemática el riego de reincidencia de las mujeres, haciendo que se les aplicaran decisiones judiciales más duras de lo requerido.

[30] HACKER, P., *op. cit. n.* 25, pp. 1148-1149.

que conducen coches rojos son más propensas a estar envueltas en accidentes, y los coches rojos son principalmente conducidos por hombres, ello implicará que los hombres, en general, acaben pagando primas más altas que las mujeres, a pesar de que el atributo género no sea utilizado de forma directa.

Es decir, por mucho que un determinado factor sea directamente excluido de un modelo, ese mismo factor acabará saliendo a la luz por su correlación con otros factores que sí se recogerán, por no ser considerados sensibles o merecedores de tal protección[31] . Otro de los ejemplos más típicos es el de la correlación del atributo protegido "raza" con el atributo relativo al código postal o lugar de residencia[32] . Es habitual la concentración de personas de un mismo origen étnico en determinados barrios, por lo que, aunque no se use el factor étnico, por su fuerte correlación con otro factor que actúa como *proxy* (indicador), la misma información acabará siendo aprehendida del modelo. Lo mismo ocurre con el sexo, pues, al examinar el algoritmo de contratación de personal que excluía los CV de mujeres[33] (¡sin que el modelo hubiera usado el factor "género" como variable!), se detectó que el modelo utilizó *proxies* que correlacionaban de forma fuerte con el género; aquellas personas en cuyos *currícula* figuraban palabras relacionadas con lo femenino (escuelas femeninas, deportes típicamente femeninos etc.), eran excluidas.

Por tanto, si bien la noción de justicia mediante la no utilización de factores protegidos encuentra apoyo dentro de la comunidad jurídica, debemos tener presentes todas las reservas (y evidencias) que la comunidad científica nos presenta.

2.3.2 Paridad estadística (también llamada equidad de grupo)

Según Corvett-Davies y Goel, implica que alguna de las medidas de error es igual para todos los grupos, definidos por sus atributos protegidos. Continúan explicitando estas medidas, que incluyen: "*la tasa de falsos positivos, la tasa de falsos negativos, la precisión, la sensibilidad, y la proporción de decisiones que son positivas. Nosotros incluimos también el área debajo de la curva ROC"* [34] .

La paridad estadística es una de las medidas que más acogida parece haber tenido en los últimos años[35] , y que ha estado en discusión, precisamente, en el caso más paradigmático

---

[31] KIM, P. T., *op. cit*. n. *27*, p. 193.

[32] CALDERS, T., & ŽLIOBAITĖ, I., *op. cit. n*. 13, p. 47. Señalan estos mismos autores que las afirmaciones sobre la necesidad de eliminar atributos sensibles y todos sus indicadores no son acertadas, o al menos no en todo caso, pues en ocasiones algunos indicadores o proxies pueden contener información relevante para la tarea predictiva o clasificatoria. El ejemplo que ofrecen es el de los códigos postales pues, aunque pueden contener indirectamente información sobre la etnicidad de la persona, pueden ser útiles si de lo que se trata es de conceder un préstamo hipotecario, donde la información sobre localización de la residencia es importante.

[33] DASTIN, J.: "Amazon scraps secret AI recruiting tool that showed bias against women", en *Reuters, 11* October 2018, enlace: https://perma.cc/328A-UJFM (consultado por última vez el 9/11/2022).

[34] CORBETT-DAVIES, S., & GOEL, S.: "The measure and mismeasure of fairness: A critical review of fair machine learning", en *arXiv preprint arXiv:1808.00023*, 2018, p. 6. En todo caso, aunque los términos a los que se alude puedan parecer extraños al jurista, se explicará a continuación qué implica cada uno de ellos de la manera más sencilla posible.

[35] Especialmente después de la publicación del estudio de la agencia Propublica; ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L.: "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks", en *ProPublica*, [en línea], 23 de mayo de 2016, disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (consultado por última vez el 9/11/2022), en el que se denuncia la falta igualdad en las tasas de falsos positivos en un conocido software de predicción de la reincidencia.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

de escándalo algorítmico, en el que una acusación de sesgo algorítmico en un programa de predicción de la peligrosidad puso sobre la mesa el debate sobre los sesgos algorítmicos que hoy ocupa el presente artículo[36] .

Aunque dentro de la paridad estadística pueden diferenciarse distintas métricas o parámetros, todos ellos tienen un denominador común: se considera que un modelo justo debe producir un error igual entre todos los grupos. Los indicadores de error son variados, y para entender algo más sobre ellos se hace necesario adentrarse, brevemente, en la lógica estadística de dichos instrumentos.

Cuando un modelo estadístico se somete a evaluación podemos encontrar diferentes medidas tendentes a examinar su actuación clasificatoria. Las más conocidas se derivan de lo que se llama "matriz de confusión", en la que se confrontan directamente las predicciones realizadas por el modelo con los resultados verdaderos[37] . Por ejemplo, centrándonos en el campo penal, la matriz compararía si las personas que el modelo consideró como propensos a reincidir realmente lo hicieron o no, y si las personas que consideró no propensas a cometer un delito se ajustaron a esa predicción[38] . Surgen así los siguientes conceptos:

- Verdaderos positivos: el modelo predijo que reincidirían y lo hicieron
- Verdaderos negativos: el modelo predijo que no reincidirían y no lo hicieron
- Falsos positivos: el modelo predijo que reincidirían, pero no lo hicieron (error tipo 1)
- Falsos negativos: el modelo predijo que no reincidirían, pero sí lo hicieron (error de tipo 2)

De dicha matriz derivan algunos valores y métricas:

- Tasa de verdaderos positivos (también llamada sensibilidad o *recall*): fracción de ejemplos predichos de forma positiva y correcta por el modelo
- Tasa de falsos positivos: fracción de todos aquellos casos clasificados como positivos sin serlo
- Tasa de falsos negativos: fracción de todos aquellos clasificados como negativos sin serlo
- Tasa de verdaderos negativos (o especificidad): fracción de casos negativos que son correctamente clasificados
- Valor de predicción positivo: probabilidad de obtener una predicción positiva siendo positivo
- Valor de predicción negativo: probabilidad de obtener una predicción negativa siendo negativo
- Área bajo la curva ROC: se trata de "*un índice global de discriminación, igual a la probabilidad de que un individuo violento seleccionado al azar reciba una clasificación de riesgo superior […]que un individuo no violento seleccionado al azar*"[39]

Una vez aclarados estos conceptos, veamos cómo influyen en la noción de justicia calificada como "paridad estadística". La paridad estadística requiere que el modelo alcance

[36] Se trata del conocidísimo caso COMPAS que se explicará con mayor detalle en el apartado 5 de este artículo.

[37] Así lo explican CORBETT-DAVIES y GOEL, op. cit, n. 34, p. 6: *"incluimos en esta definición [la de paridad estadística] cualquier medida que pueda ser computada a partir de una matriz de confusión de dos por dos, donde se representa en una tabla la distribución conjunta de las decisiones y los resultados para cada grupo"*.

[38] Una magistral explicación sobre los diferentes indicadores de capacidad predictiva y el significado que tiene usar uno u otro parámetro puede encontrarse en: MARTÍNEZ GARAY, L., MONTES SUAY, F *op. cit. n. 4*.

una proporción igualada de alguna de estas medidas de error para todos los grupos protegidos. Se trataría, por ejemplo, de que el modelo arroje una tasa de falsos positivos igual para los acusados blancos que para los negros. En otras palabras, que el modelo clasifique erróneamente como peligrosos en el mismo porcentaje en los dos grupos protegidos.

Aunque esta sea una de las principales métricas usadas para ilustrar la paridad estadística, lo cierto es que todas las métricas anteriormente referenciadas pueden servir de base para esta evaluación "paritaria"; se trata de que el error o el parámetro de corrección (sea el que sea) sea igual en todos los grupos.

Algunos autores son, sin embargo, críticos con esta noción de equidad, pues en ciertas ocasiones requiere realizar algunos ajustes sobre los datos que puede conducir a resultados injustos. Así, Goel, Shroff, Skeem y Slobogin[40] argumentan que, aunque parezca contraintuitivo, las diferencias en las tasas de falsos positivos a menudo nos dicen más sobre las poblaciones subyacentes que sobre el algoritmo mismo. De hecho, la tasa de falsos positivos puede aumentar automáticamente con la tasa global de reincidencia de un grupo, y que, inevitablemente, este patrón se mantendría incluso si la valoración del riesgo se basara simplemente en el juicio humano, y no en un modelo estadístico, pues esta elevación de la tasa de falsos positivos es una consecuencia esperable de cualquier algoritmo que calcule de forma precisa el riesgo individual.

También son críticos Kearns y Roth[41] cuando explican que la paridad estadística por sí sola no sirve para especificar el objetivo de predicción del modelo, sino que simplemente funciona como límite a esas predicciones, de tal forma que un mal algoritmo (por ejemplo, uno que eligiera los atributos predictivos de manera completamente aleatoria) podría cumplir con las exigencias de paridad estadística, pero eso no significaría que fuera un algoritmo óptimo para el objetivo predictivo. Así, continúan, este fenómeno estadístico general afecta a prácticamente cada medida de precisión. Y, como resultado, examinar las diferencias de errores entre los grupos es una medida pobre para evaluar la justicia. Asimismo, afirman estos mismos autores, exigir que las tasas de error sean iguales puede llevar, por sí mismo, a una decisión discriminatoria, pues para alcanzar esta paridad en ocasiones es necesario clasificar de forma incorrecta a individuos de bajo riesgo como de alto, y viceversa, dañando potencialmente a miembros de todos los grupos en este proceso.

2.3.3 Calibración (también denominada equidad o justicia individual)

En este caso, la medida que sirve para afirmar la corrección del modelo tiene que ver con que se alcance la misma capacidad predictiva con independencia de la pertenencia de un individuo a un determinado grupo; es decir, si un individuo es calificado con una puntuación de 8 en el umbral de riesgo, se exige que dentro de ese umbral reincidan en la misma proporción todas las personas, con independencia de si son pertenecientes a un grupo protegido o no. Trasladado, una vez más, a nuestro ejemplo: dentro del umbral de peligro 8, acusados negros y blancos tendrán la misma probabilidad de reincidir.

---

[39] SINGH, JAY P.: "Predictive validity performance indicators in violence risk assessment: a methodological primer", en *Behavioural Sciences and the Law* (31), 2013, p. 15.

[40] GOEL, S., SHROFF, R., SKEEM, J., & SLOBOGIN, C., *op. cit*. *n. 17*, pp. 16-17.

[41] KEARNS M., & ROTH, A., *op. cit*. n. 23, p. 70.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

En este sentido, dicho parámetro o noción de justicia podría encaminarse a un ideal de justicia individual, pues aboga por no tener en cuenta la pertenencia de un individuo al colectivo protegido de que se trate, con todos los problemas y contradicciones patentes que esto genera, en una realidad donde la pertenencia a colectivos vulnerables o protegidos se toma cada vez más en consideración, como más adelante se expondrá.

2.4 Valoración global

Cada una de las nociones formales de igualdad aquí propuestas presentan sus ventajas e inconvenientes. Todas ellas reflejan de algún modo algún tipo de noción de lo justo, pues se asocian a ideas que están instauradas en nuestro imaginario sobre lo que debemos entender por resultado igualitario. Así, o bien no tener en cuenta aquellos factores que no tengan relevancia a la hora de hacer una determinada clasificación, o bien intentar igualar los errores a nivel grupal, o bien intentar igualar la capacidad de predicción atendiendo a la individualidad, y no a las características grupales.

Pero se hace necesario, sin embargo, realizar una importante matización, de la que advierten con vehemencia los matemáticos y científicos de datos; dichas nociones de justicia formales son incompatibles entre sí. Aunque se prescindirá aquí de una explicación matemática, se sintetizará la problemática que ponen de manifiesto los números, con su innegable veracidad. Cuando la distribución de riesgos es diferente entre los grupos, es imposible matemáticamente satisfacer la paridad estadística y la calibración simultáneamente[42] .

Así, como señala Kim[43] , hay un desacuerdo sobre los significados de la discriminación y esta falta de consenso se ve agravada por el hecho de que puede que no sea posible satisfacer diferentes definiciones de justicia simultáneamente. Se pone el ejemplo de un sistema que trata de predecir las violaciones de la libertad condicional, de forma tal que sea justa e igualitaria para todos los grupos. Como se ha visto, la exigencia de no discriminación puede ser definida como la necesidad de igualdad la proporción de corrección en los positivos y los negativos para cada grupo pero, alternativamente, también puede ser definida como la necesidad de igualar la proporción de falsos positivos o de falsos negativos para todos los grupos. Resulta que allí donde las tasas base sean diferentes entre los grupos (por ejemplo, que una proporción más alta de hombres viole la libertad condicional que las mujeres), entonces estas tres nociones de igualdad serán incompatibles, de forma que satisfacer una de ellas hará que las otras dos no se puedan cumplir.

Es decir, cuando las tasas base reales son diferentes, (porque, por ejemplo, sea cierto que las mujeres y los hombres reinciden en proporciones diferentes), no es posible satisfacer la paridad estadística (conseguir que el modelo falle en la misma proporción cuando clasifica a hombres que cuando clasifica a mujeres), sin violar a la vez la calibración (es decir, que una mujer y un hombre que obtienen la misma puntuación de riesgo tengan la misma probabilidad de reincidir). En palabras de Solar Cayón[44] : "*cuando la distribución real del*

[42] CHOULDECHOVA, A.: "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments", en *arXiv:1703.00056*, 2017.

[43] KIM, P. T., *op. cit*. n. 27, p. 194.

[44] SOLAR CAYÓN J.I., "Inteligencia artificial en la justicia penal: los sistemas algorítmicos de evaluación de riesgos", en *SOLAR CAYÓN, J.I. (Ed.), Dimensiones éticas y jurídicas de la inteligencia artificial en el marco del Estado de Derecho*, Universidad de Alcalá, 2020, nota a pie de pp. 162-162.

riesgo de reincidencia varía entre los grupos, las diferencias en las tasas de error de grupo son el resultado lógico de un algoritmo que capture correctamente el riesgo de cada individuo: es difícil conseguir una simetría en las tasas de error de diferentes grupos cuando sus tasas base difieren significativamente". Este patrón estadístico es reconocido con el nombre de "infra-marginalidad".

Es posible variar los parámetros del modelo para obtener uno u otro resultado, dando más peso a un factor que a otro, pero lo cierto es que, en ocasiones, esto no dará lugar a un resultado en que se cumplan todas y cada una de las nociones de justicia.

Parece que llegamos, pues, a un callejón sin salida. Los científicos de datos han trabajado como una constante en el desarrollo de herramientas tecnológicas que son capaces de detectar y mitigar algunos sesgos, y es uno de los campos donde mayores investigaciones se están realizando, proponiendo nuevas métricas que tienen en cuenta uno u otro parámetro[45] . Sin embargo, no debemos olvidar la importancia que una decisión técnica como calibrar una tasa de error puede implicar. Optar por una u otra noción de justicia no puede ser una cuestión que se decida, sin mayor debate y reflexión, en los laboratorios de los desarrolladores de las más sofisticadas herramientas de *de-biasing*. Es aquí donde toma realmente importancia un término que no deja de repetirse cuando se habla de desarrollo de las nuevas tecnologías; "interdisciplinariedad". Resulta clave, pues, ahondar en la reflexión que otras disciplinas teóricas puedan aportar. Este será el objetivo del siguiente apartado.

3. LA NECESARIA PARTE SOCIAL

Regresando una vez más a ese enfoque "sociotécnico" al que se aludía en el encabezado, debemos necesariamente revisitar la literatura que pone el foco en la complejidad y variabilidad que la faceta social y contextual de este fenómeno aporta.

Selbst es claro al respecto cuando analiza los posibles fallos que puede suponer aplicar este tipo de herramientas en contextos sociales complejos. Uno de los errores por él apuntados es lo que denomina el *formalism trap*, que podríamos traducir como "trampa del formalismo", y que implica un "*fallo a la hora de tener en cuenta el significado completo de conceptos sociales tales como justicia, que pueden ser procedimentales, contextuales y discutibles, y que no pueden ser resueltos mediante formalismos matemáticos*". Hace hincapié, asimismo, en una idea fundamental; la relativa a que "*los valores normativos contenidos en el contexto social relevante son los determinantes*"[46] .

Insisten en esta idea, asimismo, Aizenberg y Van den Hoven[47] , al afirmar que estas aproximaciones han abierto el foco del diseño para incluir no solo requerimientos tecnológicos, sino también una consideración del contexto social en el que la tecnología es

[45] Véase, a título de ejemplo, la herramienta lanzada por la compañía tecnológica IBM, donde se explica la dinámica de AI Fairness 360, un *toolki*t de código abierto cuyo principal objetivo es ayudar a facilitar la transición a algoritmos justos para usar en el escenario industrial y así promover un cuadro común para los investigadores y evaluadores de algoritmos. Se trata, con esta herramienta, de promover un entendimiento profundo de las métricas de justicia y de las técnicas de mitigación, para permitir una plataforma común a los investigadores y a la industria y compartir un marco que ayude a transicionar a un marco de investigación de algoritmos justos de uso industrial. Véase: BELLAMY, R. K., DEY, K., HIND, M., HOFFMAN, S. C., HOUDE, S., KANNAN, K., ... & ZHANG, Y.: "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias", disponible enhttps://arxiv.org/abs/1810.01943, 2018, p. 1.

[46] SELBST, A. D., BOYD, D., FRIEDLER, S. A., VENKATASUBRAMANIAN, S., & VERTESI, J., *op. cit. n.* 10, p. 61 [traducción de la autora]

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

integrada, y cómo las necesidades sociales y valores deben ser traducidos a exigencias de diseño socio tecnológicas.

Así lo afirma Hacker[48] cuando señala que las decisiones normativas difíciles son inevitables. Aunque diferentes criterios de justicia han sido propuestos por la literatura, bajo circunstancias normales algunas de ellas son incompatibles, así que no queda otro remedio que decidir entendiendo los conceptos en las que se basa cada una de ellas. En la misma idea insisten Veale y Binns: "*los investigadores y los profesionales que trabajan en pro de un aprendizaje automático más justo deben reconocer que no se trata sólo de un problema abstracto de optimización restringida. Es un problema desordenado, contextualizado y necesariamente sociotécnico, y debe ser tratado como tal*"[49] .

En definitiva, la literatura advierte de que las peculiaridades contextuales (donde conviven, sin duda, factores históricos, políticos, sociales y morales), deben ser evaluadas y muy tenidas en cuenta para construir modelos que sean capaces de aportar soluciones justas en un contexto específico. En palabras de Borgesius: "*probablemente no sea útil adoptar normas para la toma de decisiones algorítmicas en general. Al igual que no adoptamos, ni pudimos, un estatuto para regular la revolución industrial, no podemos adoptar un estatuto para regular la toma de decisiones algorítmica. Para mitigar los problemas causados por la revolución industrial, necesitábamos diferentes leyes para la seguridad laboral, la protección del consumidor, el medio ambiente, etc. En los distintos sectores, los riesgos son diferentes y están en juego distintas normas y valores. Por lo tanto, las nuevas normas para la toma de decisiones algorítmicas deben ser específicas para cada sector*"[50] .

El concepto de lo justo no puede ser igual cuando se habla de distribución de recursos escasos (ej.: acceso al empleo o a la educación), cuando se trata de prestar servicios (publicidad que se recibe o acceso a préstamos), cuando se distribuyen derechos políticos (ej.: libertad de prensa o libertad de expresión), o cuando se distribuyen males (que es lo que, en última instancia, hacen las herramientas de predicción de la peligrosidad en el campo penal). La afectación a los derechos no es la misma, las implicaciones y los principios en juego no funcionan igual, y la respuesta no puede ni debe ser uniforme.

Se requieren, por tanto, soluciones específicas para contextos específicos. Las estructuras sociales complejas (como puede ser, entre otras, el sistema penal en su conjunto), responden a una serie de mecanismos, principios rectores y realidades operativas, que no deben desconocerse de cara a afrontar los problemas de justicia.

Así lo ponen de manifiesto autoras como Lin y Chen[51] , cuando proclaman que las desigualdades estructurales consecuencia de opresiones e injusticias históricas que sufren los colectivos tradicionalmente vulnerados deben ser abordadas y tenidas en cuenta, influyendo

---

[47] AIZENBERG, E., & VAN DEN HOVEN, J., *op. cit*. n. 9, p.2, hablan de una propuesta centrada en el denominado "*design for values*", donde se exige una traslación explícita de los valores sociales y morales a los requerimientos de diseño en función del contexto específico.

[48] HACKER, P., *op. cit. n*. 25, p. 33.

[49] VEALE, M., & BINNS, *op. cit. n. 19*, p. 13 [traducción de la autora].

[50] ZUIDERVEEN BORGESIUS, F. J.: "Strengthening legal protection against discrimination by algorithms and artificial intelligence" en *The International Journal of Human Rights*, vol. 24, nº 10, p. 1585. DOI: 10.1080/13642987.2020.1743976 [traducción de la autora] Véase, también, para una concepción muy similar en lo relativo a las diferentes esferas de justicia distributiva, en términos mucho más genéricos: WALZER, M., Las esferas de la justicia. Una defensa del pluralismo y la igualdad", FdE, Ciudad de México, 2016

[51] LIN, T. A., & CHEN, P. H. C., *op. cit. n. 21* .

de forma directa en toda la disciplina que se encargue de abordar los problemas de justicia de los modelos algorítmicos. Su propuesta parte de una oposición a las aproximaciones dominantes, tradicionalmente basadas en la localización de sesgos algorítmicos para asegurar cierta paridad en algunas medidas estadísticas entre determinados grupos de personas. Critican, así, las respuestas tecnocéntricas, pues conciben los sistemas de IA como entidades aisladas de los contextos sociales en que se sitúan.

Así, podemos reorientar la pregunta, para formularla de la siguiente manera: ¿teniendo en cuenta las limitaciones técnicas y el concreto contexto en el que se van a implantar estos modelos, qué concepto de lo "justo" debe prevalecer? Se examinarán, a continuación, algunas de las propuestas que, apegadas a marcos teóricos preexistentes, ponen el foco en dichos conceptos de justicia como posibles entendimientos a aplicar en determinados contextos.

Se advierte de que no se pretende, en absoluto, examinar las numerosísimas posibilidades que, desde la filosofía política, pueden acogerse como definiciones de justicia. El objetivo de este apartado es más modesto, y pretende servir como ilustración del trabajo que, en connivencia, filósofos, científicos y juristas vienen realizando cuando han pretendido aunar los conceptos más clásicos de justicia con los problemas propios de la sociedad algorítmica. Examinaremos, pues, algunas de las propuestas en que filosofía y tecnología aparecen ya unidas, pues, se acoja o no alguna de las mismas, pueden servir como punto de partida al debate que, sin lugar a duda, debe producirse en cada concreto ámbito de aplicación de las herramientas algorítmicas.

3.1 Meritocracia frente a igualitarismo

Como punto de partida, podemos coincidir en la visión aportada por Friedler, Scheidegger y Venkatasubramanian[52] cuando afrontan de forma cruda la realidad de las visiones en irremediable conflicto. Así, afirman, los investigadores deben ser explícitos sobre sus visiones sobre el mundo y su asunción de valores, pues el sistema que diseñen siempre va a codificar algún tipo de creencia. Así, para estos autores existen dos grandes creencias axiomáticas sobre el mundo, que condicionan de manera definitiva las definiciones de justicia.

- La asunción "*what you see is what you get*", que implica que los datos observables reflejan el mundo tal y como es, y que si la realidad refleja desigualdades es porque las mismas son un reflejo de la realidad y de las cualidades o elecciones de las personas envueltas en esas dinámicas.
- La asunción del "*we are all equal*"; que afirma que lo único que nos hace diferentes son las estructuras de desigualdad y poder que funcionan para oprimir a determinados colectivos, pero, por lo demás, todos somos iguales, y por ello, es necesario implementar mecanismos de corrección de esas desigualdades del tipo igualación de resultados grupales (lo que más arriba se ha definido como "justicia grupal" o "paridad estadística").

Partir de una u otra visión del mundo condicionará, sobremanera, la aceptación de los algoritmos en un determinado contexto[53] .

[52] FRIEDLER, S. A., SCHEIDEGGER, C., & VENKATASUBRAMANIAN, S.: "The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making", en *Communications of the ACM*, 64(4), 2021, pp. 136-143.

[53] Los autores no son, sin embargo, ambivalentes respecto a estas visiones. Afirman directamente que, al considerar los algoritmos para su aplicación en un determinado contexto, no todas las asunciones son igual de razonables. Así,

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

En una línea similar, aunque con otros términos, se refiere Hacker[54] a este problema. En su caso, contrasta dos visiones filosóficas tradicionales, y lo hace uniéndolas directamente con las dos vertientes de la igualdad en su significado más legalista.

- Las visiones de justicia individual se unen a la visión aristotélica de la justicia, apegada al concepto de igualdad formal. La justicia se consigue tratando a los individuos similares de manera similar. Para que esta concepción de justicia sea válida, afirma el autor, es necesario que se cumplan dos condicionamientos (uno fáctico y otro normativo)
  - Fáctico: que los datos subyacentes puedan medirse de forma adecuada (carecer de sesgos)
  - Normativo: que para el procedimiento en cuestión se considere adecuada una visión meritocrática, es decir, que los resultados estén condicionados por los logros previos y esto sea algo aceptable (se alude a logros previos como pueden ser el historial de pagos, las notas de exámenes o la tasa de accidentes).
- Las visiones de justicia grupal, apegadas a la noción de igualdad sustantiva o igualdad en los resultados, se recogen en una visión igualitaria o capabilitista de la justicia. Es decir, se parte de la idea de que los resultados que obtengan los individuos están condicionados por las distribuciones de poder y recursos injustas que se han realizado tradicionalmente.

Sin embargo, algunos autores, como Binns, han ido un paso más allá al reflexionar sobre las diferentes aristas implicadas. Así, de sus "lecciones de filosofía política[55] " podemos derivar una interesante reflexión acerca de la incapacidad del igualitarismo para hacer frente, por sí mismo, a las demandas de un ideal de justicia. Señala el autor que la cuestión de qué tipo de igualitarismo está implicado es relevante para nuestras asunciones sobre el impacto dispar de los resultados algorítmicos. Así, el autor presenta someramente alguna de las visiones "enfrentadas" dentro de las posturas igualitarias; partiendo de la pregunta ¿igualdad de qué? ("*equality of what*"). Así, señala, no puede llegarse al mismo resultado cuando de lo que se trata es de distribuir bienestar, recursos, capacidades o un estatus político o democrático. La idea que se deriva de esta apreciación es que los diferentes contextos están sujetos a diferentes esferas de justicia, y así, siguiendo el mismo ejemplo propuesto por el autor, si de lo que se trata es de asignar la capacidad de votar a los ciudadanos, no pueden entrar en consideración argumentos relacionados con la "igualdad de oportunidades", sino que solo debe entrar en juego una distribución absolutamente igual. Si, por el contrario, se trata de competir por posiciones sociales o bienes de mercado, puede que sí se deba entrar a valorar la lógica de la igualdad de oportunidades.

3.2 El igualitarismo de la suerte

Varios son los autores que han recurrido a esta concreta rama del pensamiento igualitarista para intentar encontrar una noción de la justicia apegada a conceptos que resultan, quizá, algo más familiares al Derecho penal. Esta corriente, tal y como Binns explica, tiene en cuenta el peso de la suerte bruta y de las decisiones a la hora de considerar si una

afirman, existe una extensa literatura que muestra que existen sesgos estructurales en la justicia criminal, por lo que en un campo como este no es razonable realizar una asunción del tipo "*what you see is what you get*".

[54] HACKER, P., *op*. *cit*. n. 25, p. 33 y ss.

[55] BINNS, R.: "Fairness in machine learning: Lessons from political philosophy", en *Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research* 81, 2018, pp. 1–11.

desigualdad es aceptable. Este tipo de pensadores (entre los que se insertan, entre otros, Dworkin o Arneson), diferencian entre las desigualdades que se deben al azar o a la suerte (a la que ellos denominan "suerte bruta"), de aquellas que son el resultado de opciones personales y apuestas informadas[56] .

Así, y trasladando esta teoría a la problemática propia de las herramientas de predicción del riesgo, Binns, en la obra arriba citada, afirma que la utilización de factores predictores que no dependen de la decisión libre del sujeto no es permisible. Pero, dando incluso un paso más, llega a afirmar que incluso aquellos factores que sí tienen algo de decisión, como puede ser la elección del barrio de residencia, deben "exculparse", en el sentido de que si cumplen algún tipo de propósito favorable a la sociedad (como evitar la exclusión social absoluta de estos lugares), dichas elecciones deben estar protegidas de las consecuencias negativas.

Recoge Naudts[57] también esta idea de la importancia del igualitarismo de la suerte para determinar la justicia de un resultado algorítmico. Así, lo que importa realmente es si las desigualdades voluntarias pudieran haber sido "razonablemente evitadas", por lo que, concluye: el resultado de un algoritmo será una manifestación de suerte bruta y, por tanto, injusto, cuando el resultado no sea razonablemente evitable para el individuo, o bien razonablemente previsible o bien el individuo no tuviera la habilidad (o no fuera consciente de ella) para influenciar la probabilidad del resultado.

3.3 El pensamiento de Rawls

El punto de partida de esta concepción *rawlsiana* de la justicia algorítmica tiene que ver con que algunos de los primeros científicos que iniciaron la labor de sistematizar las distintas nociones de justicia[58] , vincularan directamente el ideal de "justicia individual" como límite a la tarea clasificatoria partiendo de la idea de justicia de Rawls.

Ha sido expresamente reconocido por algunos autores (Duff[59] , Grace[60]), que la *Teoría de la Justicia* de Rawls[61] puede suponer un punto de partida adecuado para la regulación de los algoritmos de una forma respetuosa con los derechos fundamentales. Así, traslada

---

[56] BINNS, R.: *op. cit*. n. 55. No se tratarán, por la complejidad inherente a una propuesta filosófica de tanta importancia, todas las implicaciones asociadas a la corriente del "*luck egalitarianism*". Por nombrar una de las más comunes, baste dejar apuntado que uno de los principales problemas con los que se encuentra esta filosofía es la difícil demarcación entre las desigualdades que son elegidas y las que son merecidas". Véase, para mayor desarrollo del tema: DWORKIN, R.: "What is equality? Part 1: Equality of welfare", en *Philosophy & Public affairs*, 1981, pp. 185-246, y ARNESON, R.J.: "Equality and equal opportunity for welfare", en *Philosophical studies*, nº 56 (1), 1989, pp. 77-93.

[57] NAUDTS, L.: "Fair or Unfair Algorithmic Differentiation? Luck Egalitarianism As a Lens for Evaluating Algorithmic Decision-Making", *drafting paper*, 18 de Agosto de 2017, disponible en SSRN: https://ssrn.com/abstract=3043707 or http://dx.doi.org/10.2139/ssrn.3043707. El autor ofrece el siguiente ejemplo: Mary compra en una página web ciertos libros de temática filosófica, lo cual genera que el algoritmo le recomiende algunos títulos de obras filosóficas, mientras que Johny, que solo busca novelas románticas, recibe recomendaciones sobre este tipo de literatura. Esto sería un resultado desigual fruto de una decisión consciente del individuo y que, además, es previsible para el mismo. Sin embargo, además de las recomendaciones sobre literatura, el algoritmo realiza una predicción adicional; a partir de las búsquedas de novelas realiza una predicción sobre el tipo de música que los consumidores escuchan, ofreciendo descuentos para la compra de ciertos discos. Debido a que los lectores de filosofía escuchan normalmente jazz, Mary recibe descuentos para comprar discografía de este género musical, mientras Johny, por su afinidad con las novelas románticas, recibe, por su parte, descuentos para música pop. En este caso, la predicción, aunque basada en elecciones libres de los consumidores, no podría calificarse de previsible, pues Mary no tenía manera de saber que el hecho de comprar literatura filosófica le fuera a impedir recibir descuentos sobre otros tipos de estilos musicales.

[58] DWORK C., et al. *op. cit*. n. 27.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

Grace[62] los dos principios en que se basa dicha teoría de justicia al campo de la gobernanza de datos y, concretamente, a los problemas derivados de los resultados injustos que causan discriminación.

- Primer principio: cada persona tiene derecho al más amplio esquema de libertades básicas compatible con un esquema similar de libertades para el resto. La traducción de este principio para el autor implica una necesidad de que todos los ciudadanos tengan un acceso imperativo a los datos, de forma que se permita desafiar las decisiones derivadas del uso de *big data*. Es decir, es un requisito más apegado a la transparencia y la posibilidad de acceder a las posibles fuentes de desigualdad, para permitir un escrutinio igualitario de las decisiones.
- Segundo principio: Las desigualdades sociales y económicas deben organizarse de manera que se espere razonablemente que sean ventajosas para todos y estén vinculadas a puestos y cargos abiertos a todos. La traslación de este principio al mundo algorítmico implica, para Grace, que las tecnologías usadas en la toma de decisiones se usen de forma que no vuelvan a afianzar las desigualdades de poder, de riqueza y de otros recursos.

Así, el clásico principio de la diferencia, traducido al contexto algorítmico, implica que las desigualdades pueden ser aceptables siempre que conduzcan a un mayor beneficio total, beneficiando también a los más desfavorecidos. La intuición moral derivada de ello es la siguiente: que los algoritmos serán legítimos incluso cuando estén basados en datos sesgados, siempre que esto sea tenido en cuenta y mitigado en el proceso de desarrollo del modelo, y siempre que la herramienta se use para redirigir las desigualdades, no para exacerbarlas.

Nos encontramos aquí, por tanto, un igualitarismo que se centra en la clásica preocupación relativa a la distribución de recursos (típicamente de igualdad y riqueza), aunque no es ocioso pensar si, con el cambio de paradigma de las modernas sociedades de la era de la revolución tecnológica, no debiera ampliar su campo de aplicación a otras posibles parcelas donde se debe producir una distribución de recursos quizá no tan clásicos[63] .

En definitiva, se trata de explorar clásicas formas de concebir la justicia y las necesidades de distribución, con las nuevas realidades que nos rodean. Exigen, así, los nuevos tiempos,

---

[59] DUFF, A. S.: "Neo-Rawlsian Co-ordinates: Notes on A Theory of Justice for the Information Age", en *International Review of Information Ethics*, vol. 6, 12/2006, pp. 17-22.

[60] GRACE, J.: "AI Theory of Justice: Using Rawlsian Approaches to Better Legislate on Machine Learning" in Government", 29 de abril de 2020, disponible en: https://ssrn.com/abstract=3588256 or http://dx.doi.org/10.2139/ssrn.3588256.

[61] RAWLS, J.: A *Theory of Justice*, Belknap Press of Harvard University Press, Cambridge, 1971. No se tratará aquí de desarrollar el extenso e influyente pensamiento del filósofo, pero baste al lector tener en mente las ideas básicas que dan forma al razonamiento del pensador. Rawls parte de la idea de que detrás del velo de la ignorancia, en una posición original, un político con ninguna concepción previa de las desigualdades y los desequilibrios de poder, riqueza o privilegio que se derivan de factores como la clase social, la raza o factores geopolíticos, crearía políticas, que, desde esta posición de ignorancia, aseguraran un sistema de reglas justas para beneficiar a todos.

[62] Grace, *op*. *cit*. n. 60.

[63] Véase, para ilustrar otro ejemplo de cómo modernos pensadores, en el marco de la reflexión sobre las teorías de justicia a aplicar en el campo de las nuevas tecnologías, apuestan por una aproximación rawlsiana, en este caso, una apuesta por el ideal de justicia distributiva aplicada al campo de la información. VAN DEN HOVEN J. & ROOKSBY E.: "Distributive justice and the value of information: A (broadly) Rawlsian approach", en VAN DEN HOVEN J., & WECKERT J (Eds.), *Information Technology and Moral Philosophy*, Cambridge University Press, Cambridge, 2008, pp. 376–396.

una reflexión sobre qué bienes necesitamos repartir o distribuir, sobre cuáles son nuestras concepciones previas sobre dichos bienes, y apostar, decididamente, por una opción, siendo plenamente conscientes de las visiones y asunciones que dicha elección implica.

Resultará quizá, sorprendente al lector, el hecho de comprobar cómo, tras tantos siglos de debate, por suerte o por desgracia, la llegada de las tecnologías más punteras no es capaz de hacer que dejemos de lado una pregunta recurrente; ¿qué concepción de "justicia" debe primar? Cada una de las nociones formales de equidad puede emparejarse con uno de los tradicionales modelos de justicia, por lo que un análisis pormenorizado de la situación nos lleva a la temible responsabilidad de decidir.

Esta decisión no estará, esta vez, automatizada. Requiere de un debate pausado en el que se vean envueltas reflexiones con un marcado carácter ético. Quizá la atención a las capacidades sea adecuada en situaciones de distribución de recursos escasos, donde deban primarse ideales de justicia más cercanos a la "igualdad de oportunidades", pero no se podrá predicar lo mismo en los casos de distribución de derechos fundamentales de carácter político, donde el "capacitismo" y el ideal meritocrático deben dejarse a un lado para garantizar una absoluta igualdad de resultado (mismos derechos con independencia de las capacidades).

En definitiva, debemos plantear las preguntas más adecuadas a cada contexto. Regresemos, pues, al que nos es propio. Uno de los objetivos ínsitos a la misma existencia del Derecho penal es la distribución de males. El mal generado por el delito y el mal generado por la pena. En la eterna búsqueda del equilibrio entre ambos males se lleva moviendo el Derecho penal desde prácticamente el inicio de su existencia. No parece muy nuevo el debate sobre qué características deben ser tenidas en cuenta para realizar este balance; y desde luego, no es ajeno tampoco la visión de los igualitaristas de la suerte cuando aluden a términos como "decisiones tomadas con consciencia" o "suerte bruta". Sobre estos términos han corrido ríos de tinta, que no van a reproducirse aquí, pues excedería el objeto del presente trabajo. Baste solo dejar apuntado que la idea que debe guiar la reflexión es, en opinión de esta autora, una que tenga en cuenta los equilibrios y balances que en la dogmática penal juegan un papel.

Sirva este breve excurso político/filosófico como punto de partida de una necesaria reflexión conjunta, (interdisciplinar en el más puro sentido de la palabra), pues de ninguna manera las ponderaciones sobre los conceptos de justicia que envuelven a sistemas capaces de perturbar derechos de tan elevado calibre deben quedar fuera de la reflexión pausada de la dogmática penal (por supuesto, no en exclusiva, pues el papel de la filosofía y la ética es esencial en una tarea tan compleja). Algunos autores, de hecho, han comenzado ya en ese esfuerzo incipiente encaminado a encontrar los principios o guías que sean capaces de poner límites a estas herramientas. Límites que, sin duda, deben ser coherentes con el ordenamiento vigente y con los principios que inspiran nuestros modelos.

4. ¿TIENE EL DERECHO HERRAMIENTAS PARA ENFRENTAR ESTE PROBLEMA?

La idea que guiará este apartado resultará, sin duda, más familiar al jurista. Al enfrentar un campo tan complejo y rodeado de aristas peligrosas, parece adecuado tratar de encontrar alguna guía que, dentro del panorama legislativo, sea capaz de aportar alguna pista que nos ponga en un camino cierto sobre los límites que el Derecho debe marcar y cómo deben articularse las herramientas que hagan de freno de emergencia frente a las injusticias causadas por los instrumentos de predicción algorítmica.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

No resulta sencillo encontrar normativa o jurisprudencia en un campo tan novedoso como el aquí presentado, o al menos ninguna que de forma directa trate el problema de la "equidad", dejando de lado las enunciaciones genéricas y programáticas que numerosísimos instrumentos regulatorios (principalmente de *soft law*) vienen realizando en los últimos años[64] . Perseguimos como ideal y requisito indispensable sistemas justos y que no conduzcan a la vulneración de derechos fundamentales. Pero, más allá de eso, ¿qué tenemos?

Lo cierto es que la normativa que más tiene que decir al respecto de este genérico problema es, sin duda, la relativa a la no discriminación, pues el reto al que nos enfrentamos tiene que ver con tratar "injustamente" a colectivos tradicionalmente discriminados.

La normativa española, más allá de las genéricas enunciaciones que la Constitución realiza en sus artículos 14 y 9.2, venía regulando la cuestión relativa al desarrollo de la igualdad y la no discriminación de forma sectorial y muy apegada a las Directivas de la Unión Europea, cuya transposición al ordenamiento reproducía la dinámica de protección sectorial propia del entorno europeo, que a continuación se desarrollará brevemente.

Sin embargo, con la publicación en julio de 2022 de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, encontramos, quizá, algunos avances regulatorios, en una Ley que tiene vocación, según su propia Exposición de Motivos, de establecer un "*mínimo común normativo que contenga las definiciones fundamentales del derecho antidiscriminatorio español*". Se culmina, así, el proceso de transposición de las Directivas europeas en materia de discriminación, cuestión que sólo se había producido de forma parcial con la anterior Ley 62/2033, de 30 de diciembre, de medidas fiscales, administrativas y de orden social, que transponía la Directiva 2000/43/CE del Consejo, de 29 de junio de 2000, relativa a la aplicación del principio de igualdad de trato de las personas independientemente de su origen racial o étnico al ordenamiento jurídico español.

La labor más importante que viene a desarrollar esta Ley integral es la de operar como ley general, frente a la previa situación de protección sectorial, y la de cristalizar en la legislación positiva algunos conceptos que, hasta el momento, eran propios del ámbito jurisprudencial. Así, tendrán importancia, por cuestiones que en adelante se desarrollarán, conceptos como la discriminación por asociación, o la discriminación múltiple e interseccional, pues positivizan conceptos que desde la academia se reclaman para la solución de problemas actuales relacionados con la justicia algorítmica.

Resulta llamativo, y especialmente relevante en lo que a este artículo interesa, que en su artículo 3.1, apartado o), desarrollando el ámbito objetivo de aplicación de la Ley, se afirma que la misma se aplicará a la "*Inteligencia Artificial y gestión masiva de datos, así como otras esferas de análoga significación*", sentando así una base importante en cuanto se unen estos dos conceptos en una legislación específica antidiscriminación. Es el artículo 23 el encargado de desarrollar este concreto ámbito de aplicación, y lo hace con cuatro apartados que, de forma poco detallada, abordan algunos de los problemas ya apuntados, entre los que destacan: el favorecimiento por las administraciones públicas de mecanismos para que los algoritmos involucrados en la toma de decisiones tengan en cuenta criterios de minimización de sesgos, transparencia y rendición de cuentas, incluyendo evaluaciones de impacto que aborden el potencial impacto discriminatorio; la priorización de la transparencia y capacidad de interpretación de las decisiones, la promoción de la Inteligencia Artificial ética, confiable y respetuosa con los derechos fundamentales, y la promoción de un sello de calidad de los algoritmos.

[64] Véanse los instrumentos normativos *op. cit. n*. 6.

Aunque puede ser esta una base normativa interesante, pues plantea la unión de dos realidades que hasta el momento no aparecían conjuntamente, al menos no en la legislación, no es posible conocer aún, dado el poco espacio de tiempo transcurrido desde su entrada en vigor, si realmente significará un cambio en la dinámica de interpretación de las garantías de protección contra la discriminación.

Sin embargo, habida cuenta de que la legislación anterior (que bebía directamente de las Directivas comunitarias) es la que se ha venido aplicando e interpretando por los tribunales, es esta la que nos puede dar cuenta de los problemas que más frecuentemente han surgido. Por ello, se realizará un repaso de aquellos puntos de fricción de la normativa con la realidad algorítmica, en un intento por esclarecer qué mejoras son posibles, así como plantearnos si la nueva legislación está preparada para acometerlas.

Como se adelantaba, la normativa antidiscriminación que ha regido principalmente en España antes de la promulgación de la citada Ley 15/2022, de 12 de julio, era la derivada de la transposición de las Directivas europeas, que, de forma sectorial, venían a establecer un mínimo común normativo a todo el entorno europeo. Las más destacadas eran: Directiva 2000/43/CE, sobre la igualdad racial, aplicable en el empleo, el acceso a los bienes y servicios públicos y la educación; Directiva 2000/78/CE, la Directiva marco, contra la discriminación por motivos de religión o convicciones, discapacidad, edad u orientación sexual en materia de empleo, Directiva refundida sobre la igualdad de género 2006/54/CE, para los casos de discriminación en materia de empleo, Directiva 2010/41/UE, sobre la aplicación del principio de igualdad de trato entre hombres y mujeres que ejercen una actividad autónoma, por la que se deroga la Directiva 86/613/CEE del Consejo y la Directiva 2044/113/CE, sobre aplicación del principio de igualdad de trato de hombres y mujeres en el acceso a bienes y servicios y su suministro.

De esta normativa se han ido extrayendo las principales definiciones y marcos conceptuales con los que la jurisprudencia ha ido perfilando el derecho antidiscriminatorio, siempre con base en la distinción entre dos conceptos: la discriminación directa y la indirecta. Vamos a partir, por tanto, de las definiciones ofrecidas por el que, hasta julio de este año, era el instrumento normativo que transponía las definiciones europeas; la Ley 62/2033, de 30 de diciembre, de medidas fiscales, administrativas y de orden social, no solo por ser la legislación de referencia inmediatamente anterior en España, sino porque dichas definiciones son comunes en todo el entorno europeo en su conjunto. Así, la citada ley, en su artículo 28.1 apartados b) y c) establece:

> *b) Discriminación directa: cuando una persona sea tratada de manera menos favorable que otra en situación análoga por razón de origen racial o étnico, religión o convicciones, discapacidad, edad u orientación sexual.*
>
> *c) Discriminación indirecta: cuando una disposición legal o reglamentaria, una cláusula convencional o contractual, un pacto individual o una decisión unilateral, aparentemente neutros, puedan ocasionar una desventaja particular a una persona respecto de otras por razón de origen racial o étnico, religión o convicciones, discapacidad, edad u orientación sexual, siempre que objetivamente no respondan a una finalidad legítima y que los medios para la consecución de esta finalidad no sean adecuados y necesarios.*

No es baladí destacar que, como consecuencia de la aprobación de la nueva Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, las definiciones han cambiado ligeramente, pero, como es sobre el marco de la protección de mínimos ofrecido por la normativa comunitaria de aplicación directa en los Estados Miembros sobre

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

el que se han desarrollado toda la jurisprudencia y las consideraciones críticas en relación con su aplicabilidad (escasa y problemática) en el ámbito algorítmico, será este marco el que tomaremos como referencia, sin perjuicio de dejar apuntada esta diferencia, con la incertidumbre en cuanto a su aplicación que necesariamente trae con ella aparejada[65] .

Con la regulación antidiscriminación en la mano, surgen enseguida una diversidad de críticas, todas ellas prácticamente unívocas en cuanto a su escasa aplicabilidad para solucionar los problemas propios de la justicia algorítmica. A continuación, se tratará de exponer las razones que han motivado las críticas, específicamente vertidas en la pretensión de aplicar la normativa antidiscriminación a los problemas algorítmicos.

4.1 El problema de la regulación sectorial

La regulación antidiscriminación ha ofrecido tradicionalmente protección a colectivos concretos en contextos de aplicación particulares. Así, la normativa alude a las causas más típicas de discriminación, aquellas que tradicionalmente encuentran mención expresa en los textos constitucionales, sin que el fenómeno discriminatorio sea abordado como uno de tipo global.

Esto, sin embargo, y como se ha mencionado, ha sido modificado en el ámbito español por la nueva ley integral, que no solo amplía su ámbito subjetivo de protección, incluyendo colectivos no tan típicamente presentes en los textos constitucionales[66] , sino también con la ampliación del ámbito objetivo de aplicación[67] , que abarca sectores hasta ahora olvidados, entre los que destaca el de la Inteligencia Artificial, expresamente mencionado.

Se puede entender que con la promulgación de esta normativa el legislador español se hace consciente de la problemática de la sectorialidad, de la que hasta el momento había adolecido la normativa, y que había sido puesta de manifiesto en variadas ocasiones[68] , al hilo de su escasa aplicabilidad a nuevos contextos.

---

[65] Lo cierto es que merece la pena tener en cuenta esta diferencia, aunque su alcance sea aún desconocido, pues las definiciones de discriminación directa e indirecta, aunque se mantienen iguales en lo básico (discriminación directa es aquella que se produce como consecuencia de un trato desfavorable, mientras que en la indirecta el trato es aparentemente neutro), hay un elemento, que tradicionalmente había formado parte exclusivamente de la discriminación indirecta, relativo a la posibilidad de justificar dicho trato desfavorable camuflado de neutralidad, siempre que se persiguiera un fin legítimo y con unos medios proporcionados, que en la nueva legislación española sale de la definición de discriminación indirecta para formar parte de una cláusula de justificación más genérica, contenida en el artículo 2.2, donde se permite establecer diferencias de trato siempre y cuando "*los criterios para la diferenciación sean razonables y objetivos y lo que se persiga es lograr un propósito legítimo*", cambiando en cierto modo la dinámica de la justificación, que hasta ahora había sido objeto de diversas críticas por el amplio margen que otorgaba, solo en el plano de la discriminación indirecta, como más adelante se desarrollará.

[66] Véase art. 2.1: "*[…] Nadie podrá ser discriminado por razón de nacimiento, origen racial o étnico, sexo, religión, convicción u opinión, edad, discapacidad, orientación o identidad sexual, expresión de género, enfermedad o condición de salud, estado serológico y/o predisposición genética a sufrir patologías y trastornos, lengua, situación socioeconómica, o cualquier otra condición o circunstancia personal o social*".

[67] Desarrollado en el artículo 3 de la Ley 15/20022, de 12 de julio.

[68] WACHTER, S.: "Affinity profiling and discrimination by assocation in online behavioral advertising", en *Berkeley Tech. LJ*, vol. 35, 2020, p. 367. En este artículo trata el problema de los límites de la normativa europea en el concreto campo de aplicación de la publicidad online basada en el perfilado del usuario, aunque sus críticas son perfectamente válidas para abordar el genérico problema del perfilado. Se ofrece una visión completa de cómo esta normativa es claramente insuficiente para abordar los problemas de desigualdad y discriminación en un contexto en el que la discriminación por afinidad se desarrolla de formas no contempladas en los marcos normativos, por la particularidad y novedad de sus técnicas.

4.2 La práctica inaplicabilidad de la discriminación directa

La primera de las posibilidades aplicativas ante un supuesto caso de discriminación será acudir a la figura de la discriminación directa, cuyo denominador común es el trato menos favorable "por razón de" las causas de discriminación expresadas[69] . Así, explica Hacker[70] , en contextos de *machine learning*, la discriminación directa será rara. Ésta se da cuando el decisor usa explícitamente la pertenencia a un grupo protegido como *input* del modelo y asigna puntuaciones más bajas al mismo. Este tipo de discriminación no cubre, por tanto, una muy frecuente, que es la conocida como "*proxy discrimination*" (cuando se producen correlaciones a través de un criterio aparentemente neutral, que es lo que ocurre en la mayoría de los casos) y que, de hecho, como señala este autor, pueden ser fuente no solo de discriminación intencional, sino también de discriminación intencional enmascarada.

Añade el autor que todos los supuestos de sesgos producidos por errores en el procesamiento de datos (ya sea por medio del sesgo histórico o de muestreo) no ocurren, tal y como exige la ley, "por razón de" la pertenencia a ciertos grupos protegidos, por lo que la protección ofrecida por la discriminación directa resulta, en casi todos los escenarios, inaplicable.

Lo mismo afirma Wachter[71] , siendo consciente de las dificultades aplicativas de esta figura, aunque haciendo un intento por trasladar alguna variante interpretativa jurisprudencial (principalmente, la discriminación por asociación, donde no se exige que la persona que ha sufrido la discriminación sea parte del colectivo directamente protegido) para abordar este problema[72] .

Sin embargo, sí podemos encontrar alguna opinión discrepante entre los autores. Es de resaltar, por su confrontación con el resto de opiniones, la visión de Prassl, Binns, y Lyth[73] , que abogan abiertamente por aplicar la figura de la discriminación directa a los casos más típicos de discriminación algorítmica, pues consideran que acudir a la discriminación indirecta es un remedio pobre, dado el amplio conjunto de justificaciones a las que se da cabida, pudiendo conllevar lo que estos autores llaman "bucles de autojustificación[74] ". Abogan, así, por la aplicación de la discriminación directa en los casos de sesgo en la

---

[69] Merece la pena llamar la atención sobre la diferencia que, en este sentido, se aprecia respecto de la legislación antidiscriminación del ámbito de EEUU, donde aún se pone el énfasis en los motivos o intenciones de la persona para valorar los casos de discriminación, como bien se señala en: XENIDIS, R & SENDEN L.: "EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination", *en* BERNITZ U. et al. (Eds.), *General Principles of EU law and the EU Digital Order*, Kluwer Law International, 2020, p. 19.

[70] HACKER, P, *op. cit.* n. 25, pp. 9 y ss.

[71] WACHTER, S.: *op. cit.* n. 68, p. 401. La autora afirma: "*si el perfilado por afinidades no usa o infiere características protegidas como "raza" o "etnicidad" de la forma descrita en las Directivas, no se podrá reclamar por un supuesto de discriminación directa. Sin embargo, sí será posible la pretensión sobre la discriminación indirecta siempre que se pueda mostrar que las acciones basadas en afinidades o inferencias llevan a un efecto adverso y a resultados diferenciales para el grupo protegido en comparación con otros en una situación similar*".

[72] Resulta interesante, en este punto, destacar que la nueva normativa española reconoce expresamente la discriminación por asociación en su art. 6.2: "*a) Existe discriminación por asociación cuando una persona o grupo en que se integra, debido a su relación con otra sobre la que concurra alguna de las causas previstas en el apartado primero del artículo 2 de esta ley, es objeto de un trato discriminatorio*".

[73] ADAMS-PRASSL, J., BINNS, R., & KELLY-LYTH, A.: "Directly Discriminatory Algorithms", en *The Modern Law Review*, 2022, pp. 1-32.

[74] ADAMS-PRASSL, J., BINNS, R., & KELLY-LYTH, A., *op. cit. n.* 73, p. 8.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

muestra (datos no representativos), y en los casos de *proxy discrimination* [75] . A pesar de ello, los arriba citados se muestran plenamente conscientes de la difícil trazabilidad de una barrera clara entre los supuestos de discriminación directa e indirecta, así como de la escasa aplicabilidad que entre los tribunales (especialmente en el ámbito anglosajón) se realiza de la discriminación directa en los términos por ellos propuestos.

Merece la pena resaltar la conclusión con la que cierran el artículo; "*los tribunales han desarrollado una taxonomía compleja, a pesar de que el lenguaje legal de la discriminación directa es comparativamente simple. […] Alguna interpretación existente […] puede cubrir muchos casos de discriminación algorítmica, pero estas interpretaciones se desarrollaron en un contexto en que solo los humanos discriminaban, y donde la trazabilidad de los factores implicados en una decisión era, por tanto, limitada. En el contexto algorítmico, sin embargo, es […] más fácil identificar los factores que dan lugar al sesgo. […] La pregunta es, pues, si la aproximación de los tribunales para categorizar la discriminación directa adecuadamente se ocupa de los mecanismos por medio de los cuales las características protegidas configuran los resultados algoritmos*"[76] .

4.3 Los problemas asociados a la discriminación indirecta

La segunda opción sería acudir a la siguiente noción disponible en la legislación; la discriminación indirecta. Ésta exige acciones aparentemente "*neutras*", que den lugar a una "*desventaja particular*", y, en su formulación tradicional, siempre que no "*respondan a una finalidad legítima*", y siempre y cuando "*los medios para la consecución de esta finalidad no sean adecuados y necesarios*"[77] .

Estas serán normalmente las situaciones más comunes de discriminación algorítmica; que, en la dinámica que le es propia, mediante el uso de datos aparentemente neutros con la intención de obtener una decisión lo más precisa posible, causa un impacto diferente entre los distintos colectivos, dando lugar a resultados que no suponen, al menos *prima facie*, un trato menos favorable por razón de la pertenencia a un grupo protegido. Sin embargo, también esta figura encuentra problemas aplicativos, que se ponen de manifiesto al analizar pormenorizadamente cada uno de sus elementos.

---

[75] ADAMS-PRASSL, J., BINNS, R., & KELLY-LYTH, A., *op. cit*. n. 73, p. 17: "*la única pregunta es si el criterio está inevitablemente unido a la característica protegida; si no se pueden separar estas dos características el atributo protegido en cuestión del proxy tendremos una indicación de que hay discriminación directa […]*". Es más, van más allá, afirmando que daría igual que no fuera posible para un humano encontrar esta conexión "perfecta" entre factores; si el algoritmo lo hace, habrá discriminación directa. Ni siquiera consideran necesaria esta perfección total entre factores, pues tampoco lo ha pedido la jurisprudencia del TJUE (con cita del caso WABE and MH Müller Handels, ECLI:EU:C:2021:594), donde el Tribunal consideró como constitutiva de discriminación directa la prohibición de llevar "signos políticos, filosóficos y religiosos al trabajo que fueran conspicuos y de gran tamaño", por entender que este indicador estaba inevitablemente unido a manifestaciones religiosas específicas, a pesar de que el total de las personas afectadas por la norma no fueran mujeres musulmanas.

[76] ADAMS-PRASSL, J., BINNS, R., & KELLY-LYTH, A., *op. cit*. n. 73, p. 24 [traducción de la autora].

[77] Se destaca aquí, una vez más, lo puesto de manifiesto con anterioridad; la realidad de la normativa europea y de la jurisprudencia existente hasta el momento enmarcaba la posibilidad de justificación en el ámbito de la discriminación indirecta, si bien la actual legislación española se desmarca en este punto, con la consiguiente incertidumbre acerca de la interpretación conjunta de ambas normativas, aún pendiente de verificar.

a) Desventaja particular

Se trata de demostrar, en este punto, que la concreta persona ha sufrido una desventaja "particular", interpretada en el sentido de que afecta de manera desproporcionada al individuo en comparación con el grupo confrontado.

Más allá de los argumentos relacionados con la dificultad de prueba consecuencia de la poca transparencia de estas herramientas (tema aquí no abordado pero especialmente importante en este ámbito), surge una polémica más evidente: pese a que ha habido variados pronunciamientos al respecto del TJUE, no se consigue llegar a un consenso sobre en qué momento entender que existe una afectación "desproporcionada" constitutiva de esa desventaja particular. Es decir, ¿debe existir un umbral de desigualdad fijo a partir del cual se considere que la diferencia es, automáticamente, desproporcionada, dando lugar a esta desventaja exigida por la ley? Podemos (y debemos) tratar de encontrar algunos parámetros que aporten algo de seguridad. ¿Constituye una desventaja particular que se haya producido una diferencia del 80% entre mujeres y hombres? ¿y del 70%? Es esta una pregunta que no ha obtenido una solución certeza en la jurisprudencia europea, cuyas respuestas han sido variables y no determinantes en este sentido[78] .

b) Que no haya una finalidad legítima que justifique el resultado dispar

El presente problema es común a todos los casos de discriminación indirecta genérica; la necesidad de encontrar un criterio que nos permita saber, de antemano, qué causas de justificación son legítimas. Pero este genérico problema se ve acrecentado, una vez más, en el contexto algorítmico[79] .

Señala Wachter[80] la inconsistencia de la jurisprudencia del TJUE a la hora de aceptar justificaciones, aunque sí afirma que es posible atisbar algún tipo de preferencia hacia los objetivos de carácter social en general, más que intereses puramente económicos.

Así, podríamos decir, la prevención del crimen, la mejor asignación de recursos de una manera eficiente, o el ahorro de costes personales y económicos pueden ser invocados como finalidades legítimas, con no demasiado esfuerzo argumentativo. Y, en el concreto campo algorítmico, además, será fácil encontrar esta justificación, pues normalmente ocurrirá que aludiendo a un genérico fin que remotamente sirva a alguna utilidad social o económica, pueda entenderse que el requisito de la finalidad legítima se ve cumplido. Hacker[81] se muestra

---

[78] Así, en el caso resuelto por el TJUE en el asunto Isabel Elbal Moreno contra Instituto Nacional de la Seguridad Social ECLI:EU:C:2012:746, se determinó en relación con un supuesto de trabajo a tiempo parcial que, si el porcentaje de personas afectadas eran en un 80% mujeres, la acción era constitutiva de discriminación. Sin embargo, en el caso R (*Seymour-Smith*) *v Secretary of State for Employment*, ex parte Seymour-Smith, 1999 ECR I-666, también en materia de derechos laborales, se estableció que una medida que afectaba al 77.4% de los hombres y al 68.9% de las mujeres no constituía discriminación indirecta en contra de las mujeres.

[79] Se hace referencia en el apartado de la discriminación indirecta a los requisitos relativos a la justificación, pues es el lugar sistemático que tradicionalmente han ocupado en la jurisprudencia europea, que es la que se ha utilizado como base para el análisis, así como la que usan los autores aquí citados, pero los mismos argumentos podrían trasladarse si se tiene en cuenta, como lo hace nuestra actual normativa, como causa genérica de justificación, desdibujando así la diferencia entre discriminación directa e indirecta.

[80] WACHTER, S., *op. cit*. n. 68, p. 410.

[81] HACKER, P, *op. cit*. n. 25, p. 11.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

preocupado especialmente por esta cuestión cuando afirma que uno de los problemas más acusados de la poca adecuación de la legislación antidiscriminación europea es la posibilidad de que los modelos de predicción pueden sostener una "justificación fácil", basada en la pretendida precisión predictiva. Tampoco para servirnos, pues, este filtro valorativo.

c) Adecuación y necesidad del medio empleado

El siguiente escalón a valorar será, pues, el definitivo. Nos encontramos aquí con la exigencia de un test de proporcionalidad, que obliga a realizar una ponderación que tenga en cuenta si, habida cuenta del beneficio que este fin legítimo aporta, el medio empleado (en este caso, la decisión derivada del modelo algorítmico) es adecuado al fin (sirve realmente para alcanzar la finalidad propuesta) y si es necesario (si no había otra forma menos gravosa para los derechos de ese colectivo que usar esos factores como predictores).

En el punto relativo a la adecuación del medio, Hacker[82] se muestra crítico con el criterio que maneja la jurisprudencia europea, al considerar insuficiente que se exija una "*evidencia de lo apropiado del sistema para alcanzar el fin propuesto*". Esta evidencia, según este autor, es fácilmente cumplida en contextos de sesgos algorítmicos, pues simplemente aludiendo a la capacidad predictiva genérica del modelo, podrá decirse que es adecuado. Los estándares propuestos en este sentido por los académicos estadounidenses le parecen a este autor más coherentes, aunque exigen un escrutinio real de los factores concretos que condicionan la decisión algorítmica, pero esto es algo que Hacker considera queda fuera del espectro de posibilidades de la normativa antidiscriminación europea.

En el punto relativo a la necesidad en sentido estricto, este mismo autor hace referencia a la idea de que se debe valorar cuál es el coste de utilizar una base de datos sesgada y el coste de utilizar otra que no lo sea (o que lo sea en menor medida). En este sentido, las consideraciones que deben hacerse en el juicio de necesidad tienen que ver con la idea de compensar el coste de lograr unos datos más "limpios" y, por tanto más "justos", y la ganancia en materia de no discriminación que incurrir en este coste supone. Y así, afirma que "*solo un coste irrazonable debe exonerar a la persona encargada de la decisión de usar un modelo menos sesgado*" [83] . Concluye afirmando que en caso de que no sea posible esta base de datos no sesgada (como en ocasiones ocurrirá), entonces el juicio de proporcionalidad en sentido estricto debe centrarse en verificar si el proceso de decisión algorítmico reduce los sesgos en comparación con otros procedimientos (no algorítmicos) de decisión. Sólo en este caso, el uso de una clasificación discriminatoria debe ser considerada apropiada, pues maximiza la posición del grupo marginalizado, en correspondencia con la concepción *rawlsiana* del principio de la diferencia.

Aunque pudiera parecer que con este análisis se agotan los problemas asociados a la normativa algorítmica, lo cierto es que algunos autores han puesto de manifiesto, ya, que más allá de las dificultades propias de las categorías tradicionales, es necesario ampliar el foco para denunciar todas aquellas dificultades nacidas precisamente a la luz de los problemas inherentes a la dinámica algorítmica.

[82] HACKER, P, *op. cit*. n. 25, p. 18.

[83] HACKER, P, *op. cit*. n. 25. p. 19.

4.4 Los problemas específicos

Xenidis[84] pone de manifiesto tres problemas clave en esta materia, compartiendo, asimismo, algunas propuestas específicas para abordarlos, habida cuenta de la realidad tecnológica actual.

Los dos primeros tienen que ver con la escasez de los atributos o factores protegidos en la legislación, que no atienden a la dinámica discriminatoria expansiva de los algoritmos. Así, apunta, la discriminación algorítmica es absolutamente interseccional (pues combina constantemente factores dinámicos, lo cual dificulta, a su vez, tener en cuenta la potencialidad lesiva discriminatoria de esta combinación de factores) Por ejemplo, la discriminación sufrida por las mujeres afroamericanas, o por los hombres homosexuales de avanzada edad. Los factores protegidos por la legislación son de naturaleza discreta y estática, en contraste con el dinamismo y el surgimiento continuo de nuevos factores de discriminación. Esto es algo que, en opinión del autor, debe trasladarse a la regulación antidiscriminatoria, de forma que se acojan definiciones que sean capaces de captar estos factores. Se propone la utilización del concepto de "discriminación múltiple" y la exploración de la "interseccionalidad" en el contexto aplicativo de la normativa europea. Resulta positiva, en este sentido, la inclusión expresa en la normativa española de estas dos formas de discriminación, múltiple e interseccional[85] .

Asimismo, propone el mismo autor que se realice una interpretación amplia y compleja por los aplicadores del derecho, de forma que se tenga en cuenta el dinamismo de los patrones de discriminación, en el sentido de que, como la lógica de la predicción basada en el análisis de datos es del todo dinámica, capaz de encontrar combinaciones de correlaciones que puedan parecer inocuas pero lleven a nuevas formas de discriminación no perceptibles, no se ajusta a la realidad estática de la normativa antidiscriminación, que no es capaz de captar estas emergentes formas de discriminación. La propuesta en este punto es acudir a una interpretación amplia, tomando como base el principio general de no discriminación y las causas abiertas del art. 21 de la Carta de DDFF para enriquecer la aplicación de la normativa secundaria y sectorial.

Finalmente, se aborda un fenómeno específicamente problemático de la disciplina algorítmica, cuya importancia ha sido advertida ya en numerosas ocasiones a lo largo de este texto; la discriminación por *proxies*. En este sentido, merece la pena hacer algún apunte más.

Llama la atención Xenidis[86] sobre la naturaleza causal de la unión entre discriminación y características protegidas (se exige, en todo caso, que la discriminación se produzca "por razón de", implicando algún tipo de relación de causalidad), en comparación con la inevitable dependencia para los algoritmos en las inferencias y correlaciones. Sin embargo, el TJUE no ha dejado claro cuál es el grado de correlación exigido entre un *proxy* y un rasgo protegido para ser constitutivo de discriminación[87] . El problema se agrava cuando todo un universo de correlaciones se abre con la práctica algorítmica y no está claro cuándo el vínculo entre los datos y la categoría protegida es lo suficientemente fuerte. Lo que sí ha dejado claro el

[84] XENIDIS, R. *op*. *cit*. n. 28.

[85] Así, art. 3.2: "a) *Se produce discriminación múltiple cuando una persona es discriminada de manera simultánea o consecutiva por dos o más causas de las previstas en esta ley. b) Se produce discriminación interseccional cuando concurren o interactúan diversas causas de las previstas en esta ley, generando una forma específica de discriminación*".

[86] XENIDIS, R. *op*. *cit*. n. 28.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

Tribunal es su rechazo a que la legislación antidiscriminación sea extendida por analogía más allá de la lista de atributos específicamente protegidos.

El autor propone, para solucionar este problema, adoptar otra concepción de los atributos protegidos, y entenderlos no solo como un instrumento para reconocer identidades sociales merecedoras de protección, sino también como herramienta para capturar sistemas sociales y jerarquías que producen desventajas. En este último sentido, como vehículos para capturar adscripciones de valor, pueden funcionar como atajos para analizar el proceso de producción de injusticias sociales y así alcanzar un tratamiento adecuado de estos grupos.

Esta solución, en opinión del mismo autor, puede encontrar su reflejo en algunos pronunciamientos y doctrina del TJUE, cuando ha reconocido la figura de la "discriminación por asociación", de forma que la víctima de la discriminación no tiene por qué formar parte del colectivo directamente protegido (o, en su caso, no tiene por qué compartir esta información privada), sino que basta con que, por sus relaciones sociales e interacciones, pueda entenderse que se da este mismo perjuicio basado en los mismos motivos. El caso *CHEZ* (mujer que sin ser gitana ve reconocida su pretensión discriminatoria por vivir en un barrio con presencia mayoritaria de esta etnia), o el caso *Coleman* (madre de hijo discapacitado que es discriminada laboralmente por este motivo), son algunos de los ejemplos más paradigmáticos[88] . Se considera que este concepto de "discriminación por asociación" puede ser un acomodo legal adecuado para los casos de un algoritmo que usa datos sobre el comportamiento de las personas para realizar un perfil basado en inferencias en relación con *proxies* de un grupo protegido. Una vez más debe señalarse aquí la ventaja de que la legislación española lo reconozca de forma expresa.

Asimismo, encontrar un remedio para la discriminación por *proxies*, en opinión de Xenidis, pasa por aceptar un cuestionamiento de los límites de los grupos o atributos protegidos. Es un problema, por tanto, lejos de estar resuelto. En todo caso, y aunque estas figuras pueden suponer algún avance, hay que ser conscientes de que los jueces son los responsables de dibujar los límites de estos atributos de una forma contextual, adoptando un enfoque inclusivo y complejo de los mismos. Todos estos remedios "*deben ser invocados en nombre del principio de efectividad del derecho europeo antidiscriminatorio en el contexto de las actuales disrupciones tecnológicas*", pues "*el reto de la discriminación algorítmica requiere de una aproximación sustantiva o incluso transformadora a la igualdad*", ya que solo una "*IA centrada en la justicia social parece adecuada para evitar la propagación de viejas y nuevas formas de discriminación convertidas en outputs tecnológicos*"[89] .

Así, a la luz de este breve repaso por la realidad normativa y jurisprudencial, podemos concluir afirmando que es necesario seguir realizando avances en este sentido, pues solo siendo conscientes de la mecánica de funcionamiento de los algoritmos y sus errores, podremos ser capaces de exigir a nuestros legisladores y tribunales que adecúen los

[87] Sí que ha sido apuntado, en algunos casos, por el TJUE, qué tipo de correlaciones son válidas a estos efectos; véase, a título de ejemplo, factores que se pueden considerar ilegítimos por estar altamente correlacionados con ciertos atributos protegidos (el embarazo con el género, tratado en el caso Case C-177/88 Elisabeth Johanna Pacifica Dekker v Stichting Vormingscentrum voor Jong Volwassenen (VJV-Centrum) Plus, EU: C:1990:383) o lo contrario, atributos que no correlacionan lo suficiente (el país de nacimiento con la etnia, tratado en el caso Jyske Finans A/S contra Ligebehandlingsnævnet)

[88] SSTJUE asunto S. Coleman v Attridge Law and Steve Law, C-303/06, ECLI:EU:C:2008:415 y asunto CHEZ Razpredelenie Bulgaria AD v. Komisia za zashtita ot diskriminatsi, C-83/14, ECLI:EU:C:2015:480. Jurisprudencia cuya aplicación también considera adecuada a estos fines WACHTER, S., *op. cit*. n. 68.

[89] XENIDIS, R., *op. cit*. n. 28, p. 758 [traducción de la autora].

conceptos e instrumentos, abordando los problemas desde una perspectiva apegada a la realidad científica y social.

5. SUPERANDO EL CASO LOOMIS. APLICACIÓN DE LOS GENÉRICOS PROBLEMAS AL CAMPO PENAL

Una vez superada esta perspectiva más amplia, se hace necesario volver a aterrizar sobre el campo penal, pues conviene no perder de vista que el postulado que guía este artículo es el siguiente; se requieren soluciones específicas para cada contexto, capaces de conocer las posibilidades y limitaciones técnicas (apartado 2), éticas (apartado 3) legislativas y jurisprudenciales (apartado 4), para poder, ahora sí, aterrizar en el concreto contexto de aplicación en el que poder aplicar estas soluciones socio-tecnológicas.

Veamos, pues, qué ha ocurrido específicamente en el mucho más estrecho campo penal, concretamente, en uno de los terrenos donde mayores polémicas han surgido, a raíz de un caso paradigmático en lo que justicia algorítmica penal se refiere.

Allá donde existe un foro en el que se debata sobre Inteligencia Artificial, sesgos y decisiones algorítmicas asociadas al riesgo de peligrosidad criminal surge, una y otra vez, el mismo ejemplo. COMPAS[90] y Loomis[91] coparon portadas de medios y estudios innumerables sobre los problemas asociados a estas herramientas que, si bien estaban sobre la mesa desde sus inicios, no fueron tratados con especial atención hasta que la realidad incómoda de la perversión de los algoritmos pudiendo encarcelar a personas y haciéndolo de forma discriminatoria vino a hacer saltar todas las alarmas. Sin ánimo exhaustivo (pues son muchos los problemas de tipo procesal que surgen en el seno del procedimiento judicial consecuencia del recurso de Eric Loomis), se resumirán a continuación los hechos más relevantes.

COMPAS es una IA creada por la empresa Northpoint y utilizada por los tribunales de EEUU que, a través de un algoritmo no público, protegido por la ley de propiedad intelectual, combina una serie de datos para evaluar el riesgo de que una persona cometa un delito (de cara a decidir sobre su prisión provisional o de cara a concretar su pena y ejecución de la misma). En el caso Loomis se resolvió la apelación presentada por Eric Loomis, que alegaba violación de sus derechos constitucionales (entre ellos, derecho al debido proceso y derecho a la igualdad, por la utilización de un atributo protegido; el género, en este caso el masculino, como factor determinante en la decisión). El tribunal, sin embargo, no atendió a las demandas del apelante, dando por buena la utilización de la herramienta[92] .

Sin embargo, pese a que el caso judicial queda resuelto en los términos expuestos, lejos queda la resolución de la polémica. En 2016 se publica el estudio de la revista de investigación ProPublica que viene a sentar las bases del posterior debate de forma magistral, con un titular impactante y unas imágenes elocuentes[93] . A la derecha de la pantalla, la imagen de un hombre blanco con una media sonrisa de lado y expresión de suficiencia perceptible, a la izquierda, un hombre negro con expresión seria y mirada de derrota. El titular, al pie: "*Sesgo*

[90] *Correctional Offender Management Profiling for Alternative Sanctions*.

[91] Se hace referencia al caso judicial que llega al Tribunal Supremo de Wisconsin en 2016, resuelto por sentencia State v. Loomis. 881 N.W.2d 749 (Wis. 2016).

[92] MIRÓ LINARES, F.: "Inteligencia artificial y justicia penal: Más allá de los resultados lesivos causados por robots", en *UNED. Revista de Derecho Penal y Criminología*, 3ª Época, nº 20, 2018, pp. 108-109

[93] ANGWIN, J., LARSON, J., MATTU, S., & KIRCHNER, L., *op. cit*. n. 35.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

en las máquinas. Hay un software usado en todo el país para predecir a los futuros criminales. Y está sesgado en contra de los negros".

Detrás de este titular llamativo y esta imagen sobrecogedora se esconde un artículo de investigación serio, donde se pone de manifiesto uno de los problemas escondidos detrás de la *black box* del algoritmo: su tasa de error es desigual, pues, cuando las predicciones del algoritmo son erróneas (concretamente cuando considera que un delincuente es propenso a volver a reincidir erróneamente), falla en mayor medida al clasificar a las personas negras que a las blancas[94] . Este dato, importante y revelador, pone en alerta a la sociedad, pero ha llevado, sin duda, a confundir términos que no deben mezclarse. Muestra de la complejidad del asunto es el posterior debate académico que surge tras la publicación del citado estudio, donde se ponen de manifiesto algunos elementos a tener en cuenta para caracterizar correctamente el problema.

Es este un ejemplo perfecto para ilustrar la compleja problemática que aquí se ha venido tratando, pues ocurre que hay una confrontación entre las distintas nociones formales de igualdad en liza (expuestas en el apartado 2). Así es explicado con enorme rigor científico por estudios posteriores, que vienen a aportar algo de luz al debate, ya encarnizado, entre ProPublica, Northpoint[95] y la sociedad. De forma resumida, el argumento esgrimido por quienes confrontan las conclusiones de ProPublica es el siguiente: el algoritmo no discrimina más que otros, ni siquiera se puede decir que sus decisiones estén sesgadas en contra de un colectivo; lo que se pone de manifiesto aquí es un debate que tiene algo de matemático y algo de ético. La noción de equidad que Northpoint utiliza para evaluar la justicia de su *software* es la relativa a la calibración (o equidad individual), de forma que cuando dos acusados reciben la misma puntuación, con independencia de su origen racial, reincidirán en la misma proporción. En este sentido, COMPAS es un programa justo, pues a igual puntuación, igual tasa de reincidencia. El problema es que la noción que utiliza Propublica en su estudio es, también, correcta matemáticamente (la proporción de error es desigual), por lo que no mienten cuando dicen que el *software* está violando la paridad estadística[96] .

Recuérdese, sin embargo, lo explicado con anterioridad; cuando la distribución de los riesgos es diferente, es imposible satisfacer ambas propiedades a la vez. O bien se toma la decisión de manipular las tasas de error, haciendo que, a misma probabilidad de reincidencia, un acusado negro y otro blanco reciban diferentes puntuaciones, o bien se mantiene la misma puntuación para el mismo nivel de riesgo, pero se desequilibran los errores[97] .

---

[94] En concreto, y en el condado de Broward, Florida, la tasa de falsos positivos es dos veces más grande para los acusados negros que para los blancos; entre aquellos que no volverían a reincidir, el 31% de los acusados negros fueron clasificados como riesgo medio o alto, comparado con el 15% de acusados blancos. Véase: GOEL, S., SHROFF, R., SKEEM, J., & SLOBOGIN, C. *op. cit*. n 17, p. 16.

[95] La empresa publica su propia respuesta desmintiendo el error destacado por ProPublica. DIETERICH, W., MENDOZA, C., & BRENNAN, T.: "COMPAS risk scales: Demonstrating accuracy equity and predictive parity", en *Northpointe Inc*, nº 7(4), 2016.

[96] Respecto de las distintas nociones del concepto de equidad en juego en este caso: KLEINBERG, J.; MULLAINATHAN, S.; RAGHAVAN, M.: "Inherent Trade-Offs in the Fair Determination of Risk Scores", 2016, disponible en *arXiv:1609.05807*.

[97] GOEL, S., SHROFF, R., SKEEM, J., & SLOBOGIN, C. *op. cit*. n 17, p. 17, lo ejemplifican con referencia a los datos reales del Condado de Broward, Florida, de forma que, para igualar la tasa de falsos positivos según los datos de COMPAS en el condado de Broward, se debería clasificar a los acusados negros como peligrosos si su puntuación fuera 6 o superior, pero clasificar a los acusados blancos como peligrosos si su puntuación fuese 4 o superior.

No se pretende reabrir aquí el debate sobre la mayor o menor corrección de uno u otro criterio formal de justicia, simplemente llamar la atención sobre el hecho de que, una vez más, la discusión debe enmarcarse en estos términos. No es suficiente con afirmar que un programa viola una determinada noción formal de justicia y, así, dar el salto conceptual y afirmar que está sesgado y es discriminatorio. Es esta una visión reduccionista e insuficiente para un problema complejo, en el que muchísimas otras cuestiones están en juego.

Sólo una revisión profunda de todas ellas será capaz de llevarnos a una metodología adecuada de análisis de los problemas de la justicia algorítmica en el sistema penal, que sea comprensiva del marco y los límites que la ciencia impone, pero atenta a las necesarias constricciones legales, doctrinales y éticas. Este artículo trata de marcar un camino modesto, que empiece simplemente por la tarea de señalar aquellos elementos que deben estar sobre la mesa para iniciar un debate serio y profundo sobre un tema de tanta complejidad.

Modificar los resultados que arroja un algoritmo de predicción para evitar que haya factores que, aunque correlacionan positivamente con la posibilidad de reincidir, correlacionan también con factores a proteger, implica modificar el resultado inicial para "balancear" el resultado final. Esto implica no solo una apuesta por determinada concepción de justicia, sino que conlleva a su vez afirmar que, como a consecuencia de las desigualdades injustas del pasado los resultados no están balanceados, estamos dispuestos a sacrificar predicción por justicia[98] .

Por supuesto, para tomar una decisión sobre qué concepción de justicia queremos incentivar (o directamente obligar a que se implante) en este tipo de algoritmos de predicción, el enfoque normativo es necesario y hay que optar por una noción de justicia expresa. La siguiente pregunta será, pues: ¿cuál es la concepción normativa adecuada para que se inserten este tipo de modelos predictivos en el sistema penal? ¿es solo una la correcta a la luz de los principios y derechos en juego en el sistema de justicia penal? Deberían poder elaborarse una serie de parámetros dogmáticos que permitan a los programadores tener herramientas que guíen este proceso, pues, si en Derecho penal damos importancia a ciertos factores (la peligrosidad o la reincidencia), pero no a otros (la ideología o la moralidad), y hemos llegado al acuerdo de que otros nos son completamente indiferentes (por ejemplo, el color de la ropa que lleve el acusado), dichos factores (los que importan, los que no, y los que son indiferentes), deberían guiar el establecimiento de unas pautas uniformes a la hora de optar por un sistema normativo que informe las decisiones técnicas.

¿Qué concepciones de justicia hemos adoptado como propias en el sistema penal actual? No debe ni puede ser igual para todos los sistemas, o al menos no para los que tienen principios rectores que son completamente opuestos. Por ejemplo, no deberá ser la misma la solución en un sistema de valores en el que se entienda que el individuo debe sacrificarse por la comunidad, que en una sociedad donde se priman de forma mucho más patente los derechos individuales (como la libre determinación).

---

[98] Se ha llegado a poner de manifiesto, incluso, por algunos autores, si la manipulación de los datos para conseguir una paridad estadística no sería una forma de discriminación positiva. Así, HACKER, P., *op. cit*. n. 25, p. 32: "*El refuerzo de la equidad grupal (por medio de la paridad estadística) puede significar una acción positiva en el lenguaje de la ley antidiscriminación de la UE. La acción positiva describe un número de aproximaciones que van desde medidas de difusión hacia los grupos protegidos […] hasta la discriminación inversa (también llamada positiva). De acuerdo con las directivas antidiscriminación, el principio de igualdad de trato no impide a los Estados Miembro proveer con estas acciones positivas […]. Sin embargo, los contornos exactos de la justificación de la acción positiva son profundamente controvertidos tanto en los Estados Miembros como a nivel de la UE*". [traducción de la autora] Resulta interesante plantearse este debate, mucho más patente aun si de lo que se trata es de aplicar esta acción positiva en un campo donde se trata de decidir qué personas son merecedoras de una pena de prisión.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

Es esta una reflexión que Hacker[99] comparte en los siguientes términos: "*Las tensiones inherentes a la justicia algorítmica apuntan a la necesidad de preguntarse, y responder, cuestiones fundamentales de coherencia social e igualdad de nuevo en las sociedades occidentales asoladas por los crecientes niveles de desigualdad. Las respuestas deben ser implementadas por la ley; tienen que ser preparadas, discutidas y formuladas en un discurso social amplio. Ya que el algorithmic fairness nos permite la implementación perfeccionada de justicia individual o grupal, la concreta ratio elegida reflejará el compromiso normativo más profundo entre los valores individualistas-meritocráticos y los resultados-igualitarios*".

Pero todo ello no sin antes realizar una matización contextual, necesaria para quien, desde el campo de la dogmática penal, transita hacia un camino de arenas movedizas; algunas decisiones valorativas ya han sido tomadas en lo que a la dogmática penal se refiere. No todo vale, pues nuestro sistema sí han optado, de forma inequívoca, por algunos parámetros en detrimento de otros. Así, los principios y las máximas que guían nuestro Derecho penal no deben perderse de vista nunca a la hora de enfrentarnos a estas complejas decisiones. Así lo pone de manifestó Završnik[100] : "*Los códigos de procedimiento penal son el resultado de actos ya balanceados de una forma optimizada. Las doctrinas legales existentes, constitucionales y penales, ya encarnan decisiones sobre la fortaleza de los valores en competición, como la eficiencia o la equidad. Por ejemplo, en una respuesta liberal y democrática al crimen, es mejor dejar a 10 criminales libres que condenar a una persona inocente, esta es la prueba de fuego de los sistemas políticos democráticos o autoritarios, y los expertos en tecnología están negociando el balance de estos actos*".

Se pone de manifiesto, así, de forma explícita, que pese a que la tarea de dibujar un mapa donde entren en juego las "nuevas" nociones de equidad, dicho mapa nunca puede partir de un lienzo en blanco, donde todo sea negociable, y especialmente no puede ser así en un campo como el penal, donde los derechos en juego son fundamentales y donde, por ello, la ciencia no es libre para campar a sus anchas, dejando de lado los juicios y valores que la sociedad ha blindado. Los principios propios de la dogmática penal y los derechos fundamentales son, aquí, punto de partida inexorable para construir estos nuevos conceptos.

6. CONCLUSIONES

La lectura del artículo deje al lector, tal vez, con la desazonadora sensación de que se han planteado más interrogantes que respuestas. Retomemos, pues, las incógnitas planteadas en un inicio, para ver si algo de luz hemos conseguido arrojar a esta compleja cuestión, habida cuenta de todas las matizaciones que se han realizado a lo largo del artículo.

- **¿Es posible trasladar al lenguaje matemático conceptos como "igualdad", "equidad" o "no discriminación"?**

  La comunidad científica trata, con ahínco, en los últimos tiempos, de dar con la fórmula perfecta capaz de concentrar, de forma precisa y objetivable, todos aquellos elementos que deben formar parte de una decisión justa. La asignación de pesos específicos (por ejemplo, en forma de cuantificación exacta de los costes y beneficios que cada alternativa aporta),

[99] HACKER, P, *op. cit*. n. 25, p. 34 [traducción de la autora].

[100] ZAVRŠNIK, A.: "Algorithmic justice: Algorithms and big data in criminal justice settings", en *European Journal of criminology*, nº 18(5), 2021, pp. 629 y ss. [traducción de la autora].

ha sido propuesta, llegando a reformular el ideal de justicia en términos de funciones de utilidad[101] . Herramientas y kits de procesamiento de los datos de entrenamiento con el objetivo de detectar y mitigar sesgos salen al mercado tecnológico continuamente. Sin embargo, por mucha aproximación científica que se presente (cuyo estudio, insisto, no es en absoluto baladí), se ha mostrado otra realidad innegable; no es posible abordar la justicia de las herramientas sin tomar partido por una opción ética. Por tanto, la respuesta parece evidente. No. Las ciencias matemáticas y la computación no son capaces de programar la igualdad, como si de un concepto mecanizado se tratara.

- **¿Hay varios conceptos de lo "justo"? ¿Son compatibles? ¿Qué resultados arroja cada uno?;**

  La respuesta a esta pregunta ha sido sobradamente tratada a lo largo de estas páginas. Queda, por tanto, por realizar aquí una tarea de reflexión colectiva. Dado que los conceptos de lo justo tienen contornos difíciles de aprehender, y dado que una concepción uniforme (al menos en cada contexto), viene siendo necesaria para poder avanzar en términos de justicia y no discriminación, tal y como exigen todos los instrumentos normativos europeos, se hace absolutamente necesario profundizar en el último de los interrogantes; ¿qué resultados arroja cada concepto?, o más bien, ¿qué principios hay en juego en cada uno de los contextos sociales complejos? ¿cuáles son los principios propios del Derecho penal que debemos encontrar reflejados en ese concepto "sociotécnico" de lo justo?

- **¿Es posible tender un puente entre ambos lenguajes que brinde resultados objetivamente más justos?;**

  No solo es posible, sino absolutamente imprescindible, que este puente se levante, y debe hacerse cuanto antes. No es extraño encontrar ya algunas reflexiones en las que científicos y jurídicos entremezclan sus saberes, pero no debemos quedarnos en la superficie de este necesario intercambio de conocimiento. Si de alguna forma es posible llegar a resultados "objetivamente más justos" es precisamente así, trabajando en estrecha colaboración para que ninguna de estas dos patas esenciales quede descompensada; poco puede hacer un matemático si un jurista no le indica hasta dónde le está permitido llegar con su programación, así como también poco puede hacer un jurista cuando un matemático le dice que los límites que trata de imponer no tienen sentido desde un punto de vista científico. Dejemos, pues, que estos dos saberes se entremezclen, dando lugar a una genuina interdisciplinariedad, que se presenta también, sin duda, como piedra angular en la evolución de las nuevas tecnologías.

- **¿De qué manera el desarrollo legal y jurisprudencial de derechos como la igualdad y la no discriminación deben afectar en la programación?**

  Es nuestra labor específica como juristas la de estar bien atentos a las tendencias que, tanto en materia legislativa como en materia jurisprudencial, puedan tener algo que aportar a una disciplina novedosa pero tan inevitablemente unida a los problemas tradicionales.

El camino de una disciplina como el *algorithmic fairness* está en sus inicios. Aún tenemos la oportunidad (y la responsabilidad) de actuar colectivamente, de someter todas estas cuestiones al debate público, de que la justicia y equidad de las decisiones algorítmicas formen parte de la discusión, sometiendo estas complejas decisiones al escrutinio de la

[101] CORBETT-DAVIES, S., & GOEL, S., *op. cit*. n 34, p. 6 y ss.

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

ciudadanía, tal y como debe someterse cualquier decisión que, capaz de inmiscuirse en la esfera más íntima de su personalidad, pueda causar estragos en los derechos más fundamentales de los individuos.

Es una responsabilidad social, estructural y urgente la que esta disciplina nos plantea. Debemos ser cautos y estar atentos a todo lo que está en juego. Quizá un solo titular no sea capaz de resumir, tan magistralmente como ProPublica hizo, las complejidades y aristas que las herramientas de predicción de la peligrosidad encierran para nuestros derechos más básicos; no será una fórmula la que nos saque de este complicado atolladero, pero, en tanto seamos capaces de iniciar una discusión que tenga en cuenta todos y cada uno de los factores implicados, estaremos un paso más cerca de conseguir lo que, en realidad, todos perseguimos; una sociedad más justa, sea con algoritmos o sin ellos.

Bibliografía

ADAMS-PRASSL, J., BINNS, R., & KELLY-LYTH, A.: "Directly Discriminatory Algorithms", en *The Modern Law Review*, 2022, pp. 1-32

AIZENBERG, E., & VAN DEN HOVEN, J.: "Designing for human rights in AI", en *Big Data & Society*, July–December 2020, pp. 1-14, https://doi.org/10.1177/2053951720949566

ANDRÉS PUEYO, A., & ECHEBURÚA ODRIOZOLA, E.: "Valoración del riesgo de violencia: instrumentos disponibles e indicaciones de aplicación", en *Psicothema*, vol. 22, nº 3, 2010, pp. 403-409

ANDRÉS PUEYO, A., & REDONDO ILLESCAS, S., "Predicción de la violencia: entre la peligrosidad y la valoración del riesgo de violencia", en *Papeles del Psicólogo*, vol. 28, núm.3, septiembre-diciembre, 2007, pp. 157-173

ANDRÉS-PUEYO, A., "Peligrosidad criminal: análisis crítico de un concepto polisémico, en MAROTO CALATAYUD y DEMETRIO CRESPO (Coords.), "Neurociencias y derecho penal: nuevas perspectivas en el ámbito de la culpabilidad y tratamiento jurídico-penal de la peligrosidad", ed. Edisofer, 2013

ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L.: "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks", en *ProPublica*, [en línea], 23 de mayo de 2016, disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (consultado por última vez el 9/11/2022)

ARNESON, R.J.: "Equality and equal opportunity for welfare", en *Philosophical studies*, nº 56 (1), 1989, pp. 77-93

BABUTA, A., & OSWALD, M.: "Data analytics and algorithms in policing in England and Wales: Towards a new policy framework" en *RUSI Occasional Paper*, 2020, ISSN 2397-0286 (Online); ISSN 2397-0278 (Print).

BELLAMY, R. K., DEY, K., HIND, M., HOFFMAN, S. C., HOUDE, S., KANNAN, K., ... & ZHANG, Y.: "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias", 2018, disponible en arXiv:1810.01943

BINNS, R.: "Fairness in machine learning: Lessons from political philosophy", en *Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research* 81, 2018, pp. 1–11

BRANDARIZ GARCÍA, J. Á.: *El modelo gerencial-actuarial de penalidad: Eficiencia, riesgo y sistema penal*, Dykinson, Madrid, 2016

CALDERS, T., & ŽLIOBAITĖ, I.: "Why unbiased computational processes can lead to discriminative decision procedures" en CUSTERS B., CALDERS, T., SCHERMER, B, & ZARSKY, T. (Eds.), *Discrimination and Privacy in the Information Society*, Springer, Berlin, Heidelberg, 2013, pp. 43-57

CHOULDECHOVA, A.: "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments", 2017, en arXiv:1703.00056

COLLINS E.: "Punishing Risk", en *Georgetown Law Journal*, 107(1), 2018, pp. 57-108

CORBETT-DAVIES, S., & GOEL, S.: "The measure and mismeasure of fairness: A critical review of fair machine learning", 2018, disponible en *arXiv:1808.00023*

DASTIN J.: "Amazon scraps secret AI recruiting tool that showed bias against women", en *Reuters, 11* October 2018, enlace: https://perma.cc/328A-UJFM (consultado por última vez el 9/11/2022)

DIETERICH, W., MENDOZA, C., & BRENNAN, T.: "COMPAS risk scales: Demonstrating accuracy equity and predictive parity" en *Northpointe Inc*, nº *7*(4), 2016

DUFF, A. S.: "Neo-Rawlsian Co-ordinates: Notes on A Theory of Justice for the Information Age", en *International Review of Information Ethics*, vol. 6, nº 12, 2006, pp. 17-22

DWORK C., et al.: "Fairness Through Awareness" en Proc. *3rd Innovations Theoretical Computer Science. Conf.*, 2012

DWORKIN, R.: "What is equality? Part 1: Equality of welfare", en *Philosophy & Public affairs*, 1981, pp. 185-246

FEENBERG, A., *Transforming technology: A critical theory revisited* / (2 ed.), Nueva York, Oxford University Press, 2002

FRIEDLER, S. A., SCHEIDEGGER, C., & VENKATASUBRAMANIAN, S.: "The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making" en *Communications of the ACM, 64*(4), 2021, pp. 136-143.

FRIEDMAN, B. & NISSEMBAUM, H.: "Bias in computer systems," en *ACM Transactions on Information Systems*, July, vol.14, nº 3, 1996, pp. 330–347.

GOEL, S., SHROFF, R., SKEEM, J., & SLOBOGIN, C.: "The accuracy, equity, and jurisprudence of criminal risk assessment", en VOGEL, R. (Ed)., *Research handbook on big data law. Intellectual Property Forum: Journal of the Intellectual and Industrial Property Society of Australia and New Zealand*, ed. Edward Elgar Publishing, 2021, pp. 9-28

GRACE, J.: "AI Theory of Justice: Using Rawlsian Approaches to Better Legislate on Machine Learning in Government", 29 de abril de 2020, disponible en: https://ssrn.com/abstract=3588256 or http://dx.doi.org/10.2139/ssrn.3588256

HACKER, P.: "Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law" en *Common Market Law Review*, nº *55*(4), 2018, pp.1143–1186

HAN, *En el enjambre* (1a edición.), Herder, Barcelona, 2014

HARAWAY, D., *Manifiesto para Cyborgs: [ciencia, tecnología y feminismo socialista a finales del siglo XX]*, Buenos Aires, Letra Sudaca Ediciones, 2018

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

KEARNS M., & ROTH, A.: *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*, Oxford University Press, 2020

KEHL, D., GUO P., KESSLER S.: "Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing" en *Responsive Communities Initiative, Berkman Klein Center for Internet & Society*, Harvard Law School, 2017

KIM, P. T.: "Auditing algorithms for discrimination" en *University of Pennsylvania Law Review Online*, nº 166, 2017, pp. 189-204

KLEINBERG, J.; MULLAINATHAN, S.; RAGHAVAN, M.: "Inherent Trade-Offs in the Fair Determination of Risk Scores", 2016, disponible en *arXiv:1609.05807*

LATOUR, B., *Reensamblar lo social: una introducción a la teoría del actor-red*, Buenos Aires, Manantial, 2008

LESSIG, L., *Code: and Other Laws of Cyberspace, Version 2.0*, New York, NY: Basic Books, 2006

LIN, T.A., & CAMERON CHEN, P. H..: "Artificial Intelligence in a Structurally Unjust Society", en número futuro de *Feminist Philosophy Quarterly*, disponible en la actualidad en: https://philpapers.org/rec/LINAII-2 (consultado por última vez el 9/11/2022)

MARTÍNEZ GARAY, L., MONTES SUAY, F.: "El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias" en *InDret*, 2/2018

MIRÓ LINARES, F.: "Inteligencia artificial y justicia penal: Más allá de los resultados lesivos causados por robots", en *UNED. Revista de Derecho Penal y Criminología, 3ª Época*, nº 20, 2018, pp. 87-130

MONAHAN, J.: "A jurisprudence of risk assessment: Forecasting harm among prisoners, predators, and patients", en *Virginia Law Review*, 2006, pp. 391-435

NAUDTS, L.: "Fair or Unfair Algorithmic Differentiation? Luck Egalitarianism As a Lens for Evaluating Algorithmic Decision-Making", drafting paper, 18 de Agosto de 2017, disponible en SSRN: https://ssrn.com/abstract=3043707 or http://dx.doi.org/10.2139/ssrn.3043707

QUIJANO-SÁNCHEZ, L., LIBERATORE, F., CAMACHO-COLLADOS, J., & CAMACHO-COLLADOS, M.: "Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police", en *Knowledge-Based Systems*, nº 149, 2018, pp. 155-168. https://doi.org/10.1016/j.knosys.2018.03.010

RAWLS, J.: *A Theory of Justice*, Belknap Press of Harvard University Press, Cambridge, 1971

RIGAKOS, G.S., "Risk society and actuarial criminology: prospect for critical discourse", en *Canadian Journal of Criminology*, 41 (2), 1999, pp. 137-150.

RIVERA BEIRAS, I.: "Actuarialismo penitenciario. Su recepción en España" en *Revista Crítica Penal y Poder,* nº 9, 2015, pp.102-144

SELBST, A. D., BOYD, D., FRIEDLER, S. A., VENKATASUBRAMANIAN, S., & VERTESI, J.: "Fairness and abstraction in sociotechnical systems" en *Proceedings of the conference on fairness, accountability, and transparency*, January 2019, pp. 59-68.

SINGH, JAY P.: "Predictive validity performance indicators in violence risk assessment: a methodological primer", en *Behavioural Sciences and the Law* nº 31, 2013, pp. 8–22.

SOLAR CAYÓN J.I., "Inteligencia artificial en la justicia penal: los sistemas algorítmicos de evaluación de riesgos", en SOLAR CAYÓN, J.I. (Ed.), *Dimensiones éticas y jurídicas de la inteligencia artificial en el marco del Estado de Derecho*, Universidad de Alcalá, 2020, pp. 125-172

STEVENSON, M.: "Assessing risk assessment in action", en *Minn. L. Rev., 103*, 2018, pp. 303-384

SURESH, H., & GUTTAG, J.: "A framework for understanding sources of harm throughout the machine learning life cycle", en *Equity and access in algorithms, mechanisms, and optimization*, (*EAAMO '21*), October 5–9, 2021, pp. 1-9

VALLS PRIETO, J.: *Inteligencia artificial, Derechos Humanos y bienes jurídicos*, Thomson Reuters, Aranzadi, 2021

VAN DEN HOVEN J. & ROOKSBY E.: "Distributive justice and the value of information: A (broadly) Rawlsian approach" en VAN DEN HOVEN J., & WECKERT J (Eds.), *Information Technology and Moral Philosophy*, Cambridge University Press, Cambridge, 2008, pp. 376–396

VEALE, M., & BINNS, R.: "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data" en *Big Data & Society*, July–December, 2017, pp. 1-17

WACHTER, S.: "Affinity profiling and discrimination by assocation in online behavioral advertising" en *Berkeley Tech. LJ*, vol. *35*, 2020, pp. 368-430

WALZER, M., *Las esferas de la justicia. Una defensa del pluralismo y la igualdad*, FdE, Ciudad de México, 2016

XENIDIS, R. & SENDEN L.*: "*EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination" *en* BERNITZ U. et al. (Eds.), *General Principles of EU law and the EU Digital Order*, Kluwer Law International, 2020, pp. 151-182

XENIDIS, R.: "Tuning EU equality law to algorithmic discrimination: Three pathways to resilience" en *Maastricht Journal of European and Comparative Law*, nº *27*(6), 2020, pp. 736-758

ZAVRŠNIK, A.: "Algorithmic justice: Algorithms and big data in criminal justice settings" en *European Journal of criminology*, nº *18*(5), 2021, pp 623-642

ZLIOBAITE, I.: "Fairness-aware machine learning: a perspective", 2017, disponible en *arXiv:1708.00754*

ZUIDERVEEN BORGESIUS, F. J.: "Strengthening legal protection against discrimination by algorithms and artificial intelligence" en *The International Journal of Human Rights*, vol. 24, nº 10, pp. 2020 1572–1593, https://doi.org/10.1080/13642987.2020.1743976

Apéndices

ANEXO DE NORMATIVA

Carta de Derechos Fundamentales de la Unión Europea, aprobada en DOCE de 18 de diciembre de 2000, (2000/C 364/01)

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838

Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno, adoptado por el CEPEJ durante su 31ª Reunión plenaria, adoptada en Estrasburgo los días 3 y 4 de diciembre de 2018 (CEPEJ(2018)14)

Constitución española, aprobada en BOE núm.311, de 29 de diciembre de 1978

Convenio para la Protección de los Derechos Humanos y de las Libertades Fundamentales, hecho en Roma el 4 de noviembre de 1950, y enmendado por los Protocolos adicionales números 3 y 5, de 6 de mayo de 1963 y 20 de enero de 1966, respectivamente, ratificado por España mediante BOE nº 243, de 10 de octubre de 1979

Directiva 2000/43/CE del Consejo, de 29 de junio de 2000, relativa a la aplicación del principio de igualdad de trato de las personas independientemente de su origen racial o étnico, publicada en DOCE nº 180, de 19 de julio de 2000

Directiva 2000/78/CE del Consejo, de 27 de noviembre de 2000, relativa al establecimiento de un marco general para la igualdad de trato en el empleo y la ocupación, que protege contra la discriminación por motivos de religión o convicciones, discapacidad, edad u orientación sexual en materia de empleo, publicada en DOCE nº 303, de 2 de diciembre de 2000

Directiva 2006/54/CE del Parlamento Europeo y del Consejo, de 5 de julio de 2006, relativa a la aplicación del principio de igualdad de oportunidades e igualdad de trato entre hombres y mujeres en asuntos de empleo y ocupación (refundición), publicada en DOUE nº 204, de 26 de julio de 2006

Directiva 2010/41/UE del Parlamento Europeo y del Consejo, de 7 de julio de 2010, sobre la aplicación del principio de igualdad de trato entre hombres y mujeres que ejercen una actividad autónoma, y por la que se deroga la Directiva 86/613/CEE del Consejo, publicada en DOUE nº 180, de 15 de julio de 2010.

Directrices éticas para una IA fiable, redactado por el Grupo de expertos de alto nivel sobre inteligencia artificial (IA) constituido por la Comisión Europea en junio de 2018, en el texto publicado el 8 de abril de 2019

Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación. BOE núm. 167, de 13 de julio de 2022

Ley 63/2003, de 30 de diciembre, de medidas fiscales, administrativas y del orden social, publicada en BOE nº 313, de 31 de diciembre 2003

Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva de mujeres y hombres, publicada en BOE nº 71, de 23 de marzo de 2007

Libro Blanco sobre la Inteligencia artificial, aprobado por la Comisión el 19 de febrero de 2020, COM(2020) 65 final

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, Bruselas, 21 de abril de 2021, COM(2021) 206

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))

ANEXO DE JURISPRUDENCIA

Sentencia de la Corte Suprema de Wisconsin, de 13 de julio de 2016, asunto State v. Loomis. 881 N.W.2d 749

Sentencia del Tribunal Constitucional 103/1983, de 22 de noviembre

Sentencia del Tribunal Constitucional 128/1987, de 16 de julio

Sentencia del Tribunal Constitucional 200/2001, de 4 de octubre

Sentencia del Tribunal Constitucional 22/1981, de 2 de julio

Sentencia del Tribunal de Justicia de la Unión Europea (Gran Sala), de 16 de julio de 2015, asunto CHEZ Razpredelenie Bulgaria AD v. Komisia za zashtita ot diskriminatsi, C-83/14, ECLI:EU:C:2015:480

Sentencia del Tribunal de Justicia de la Unión Europea (Gran Sala), de 15 de julio de 2021, asunto IX contra WABE eV y MH Müller Handels GmbH contra M, C-804/18 y C-341/19, ECLI:EU:C:2021:594

Sentencia del Tribunal de Justicia de la Unión Europea de 22 de noviembre de 2021, asunto Isabel Elbal Moreno contra Instituto Nacional de la Seguridad Social y Tesorería General de la Seguridad Social, C-385/11, ECLI:EU:C:2012:746

Sentencia del Tribunal de Justicia de la Unión Europea de 6 de abril de 2017, asunto Jyske Finans A/S contra Ligebehandlingsnævnet, C-668/15, ECLI:EU:C:2017:278

Sentencia del Tribunal de Justicia de la Unión Europea de 8 de noviembre de 1990, caso Elisabeth Johanna Pacifica Dekker v Stichting Vormingscentrum voor Jong Volwassenen (VJV-Centrum) Plus, C-177/88, ECLI:EU:C:1990:383

Sentencia del Tribunal de Justicia de la Unión Europea de 9 de febrero de 1999, asunto R (*Seymour-Smith*) *v Secretary of State for Employment*, ex parte Nicole Seymour-Smith and Laura Perez, C-167/97, ECLI:EU:C:1999:60

Sentencia del Tribunal de Justicia de la Unión Europea, de 17 de julio de 2008, asunto S. Coleman v Attridge Law and Steve Law, C-303/06, ECLI:EU:C:2008:415

Estudios Penales y Criminológicos., 44(ext) (2023). ISSN-e: 2340-0080
https://doi.org/10.15304/epc.44.8838