

Derecho de la neurología: Libro blanco para el desarrollo de la
Carta de Derechos Digitales en el ámbito de la neurotecnología
Beatriz Gregoraci, Jorge Castillo Abella y Francisco Velasco Caballero (dirs.)
ISBN: 979-13-87913-93-3
Madrid, 2026
pp. 847-863
DOI: 10.37417/derecho-de-la-neurotecnologia/22
Marcial Pons Ediciones Jurídicas y Sociales
Editado bajo licencia Creative Commons Attribution 4.0 International License

CAPÍTULO XXII

EL SESGO DISCRIMINATORIO Y EL DERECHO A LA IGUALDAD EN LOS SISTEMAS DE IA

Andrea BRAVO BOLADO
Universidad Autónoma de Madrid
andrea.bravo@uam.es

1. EL PROBLEMA DE LA JUSTICIA Y LA IGUALDAD DESDE UNA PERSPECTIVA TÉCNICA. ESPECIAL ATENCIÓN A UN *LEADING CASE* MAL ENTENDIDO

En un momento en el que la IA es ya una parte indiscutible de diversos ámbitos de nuestras vidas, parece inevitable que, cuando se trata de discutir acerca de estas tecnologías y su posible impacto en la igualdad, complejos términos como «redes neuronales», «paridad estadística» o «tasa de falsos positivos» se entremezclen con términos jurídicos tales como «discriminación inversa», «colectivos vulnerables» o «factores protegidos».

Con total certeza el segundo grupo conceptual resultará más cómodo y familiar al jurista, acostumbrado a la sutileza de los matices terminológicos y a la inevitable variabilidad de los conceptos. Sin embargo, en el campo de la ciencia de datos, dichos matices no son tan comunes. Los conceptos tienen un significado bien definido, pues las fórmulas permiten lo que los conceptos, con su vaguedad intrínseca, excluyen.

Lo interesante que plantea la unión de dos mundos tradicionalmente antagónicos es, precisamente, que nos obliga a enfrentarnos a retos hasta ahora desconocidos, llegando a plantear preguntas tan estimulantes como si es posible traducir la igualdad o la equidad a una fórmula matemática¹.

¿Por qué esta necesidad de unión? El auge de la IA ha puesto ya sobre la mesa variados ejemplos de cómo un instrumento, cuando es desarrollado e implantado en marcos complejos, con interacción de las personas y sus aún más complejos problemas, ha dado lugar a situaciones de desigualdad y discriminación intolerables.

Se nombrarán, a continuación, algunos ejemplos recientes en que el desarrollo de potentes herramientas basadas en IA ha supuesto la vulneración del mandato de no discriminación. Un algoritmo de contratación creado por Amazon que no selecciona los perfiles de las candidatas mujeres. Una herramienta de detección de caras de Google que etiqueta a las personas negras como gorilas. Un *software* ofrecido por las autoridades neozelandesas para renovar pasaportes que no reconoce los ojos de una persona asiática. Una herramienta de valoración del riesgo de reincidencia que clasifica con peores resultados a los acusados negros que a los blancos.

Salta a la vista cómo, sin necesidad de entrar en los entresijos de la programación de estos sistemas, algo parece no estar bien en términos de igualdad. Para el derecho existen ciertos resultados intolerables que, por muy amparados por la ciencia que se presenten, deben someterse a examen riguroso. Pero tan compleja disciplina requiere, asimismo, un análisis complejo. Veamos, pues, dónde nacen los problemas y qué podemos hacer para solucionarlos.

1.1. Un primer paso: la necesaria toma de conciencia. La igualdad y no discriminación como criterios rectores en los instrumentos normativos de la Unión Europea

Son muchos los instrumentos normativos y de *soft law* que contemplan la igualdad y no discriminación como principios irrenunciables en el desarrollo de los nuevos instrumentos de inteligencia artificial. El ejemplo paradigmático es el del Reglamento europeo de IA (*AI Act*). Uno de sus objetivos centrales es la protección de derechos fundamentales, entre los que se encuentran, expresamente mencionados en el texto, la no discriminación y la igualdad entre hombres y mujeres, reco-

¹ Para obtener una respuesta a esta incógnita, tratando de desentrañar de forma más completa si es posible hablar de justicia en términos matemáticos, *vid.*: BRAVO BOLADO, A., (2023), «Justicia algorítmica. Un enfoque sociotécnico». *Revista de Estudios Penales y Criminológicos*, núm. 44.

nocidos en los artículos 21 y 23 de la Carta de Derechos Fundamentales de la UE, respectivamente.

Aunque *infra* se desarrollarán los aspectos relativos a la regulación sustantiva, baste con dejar por ahora apuntado que en este documento se deja entrever una preocupación patente por estos bienes, reconociendo su estatus de derechos fundamentales (dicha preocupación está patente desde las primeras versiones y borradores de este documento). Se apuesta, por tanto, por un desarrollo tecnológico donde los derechos estén en el centro.

Es suficiente por ahora señalar cómo, desde esta regulación, podemos empezar a darnos cuenta de que una nueva disciplina empieza a entremezclarse con el derecho; y no es otra que la ciencia de datos. Así, para garantizar el respeto de la igualdad entre colectivos, se comienzan a introducir conceptos que no son tan comunes cuando hablamos de discriminación en términos jurídicos. El propio Reglamento, además de mencionar los derechos tal y como aparecen reflejados en las tradicionales formulaciones de la Carta de Derechos Fundamentales, comienza a introducir nuevos términos, todos ellos llamados a poner el foco en los datos. Conceptos como «representatividad, ausencia de errores, totalidad o propiedades estadísticas adecuadas» tienen cabida en este nuevo instrumento jurídico, dando a entender que quizá la manera de evaluar la permisibilidad de determinados resultados deba alejarse de las personas y sus intenciones o actitudes (como tradicionalmente ocurre en el ámbito de la no discriminación en sentido jurídico), para centrarse en el análisis de un nuevo concepto omnipresente y casi deificado: el dato. Este relativamente nuevo concepto nos lleva a examinar cuestiones hasta el momento desconocidas.

1.2. La ciencia del dato como ciencia de la equidad

La reciente disciplina de la ciencia de datos que se ocupa de los problemas surgidos en torno al concepto de igualdad y no discriminación se denomina *algorithmic fairness*, y, aunque la traducción literal se relacione con una idea de «justicia» (*fairness*), realmente esta rama de la ciencia de datos tiene que ver con los problemas vinculados con la falta de equidad y la posible causación de discriminación, especialmente a colectivos vulnerables o que tradicionalmente han sido víctimas de discriminaciones de diversa índole.

Así, cuando se trata de estudiar el impacto que en los derechos fundamentales a la igualdad y no discriminación generan las modernas herramientas de IA, necesariamente debemos entender el proceso por el cual se generan los resultados dañosos, para entender en qué momentos o procesos deben introducirse cambios.

Cuando hablamos de resultados discriminatorios tras la implantación de un sistema de IA, no se trata de juzgar si una determinada persona de la cadena de producción, como podría ser el programador o el implementador, ha manifestado una determinada actitud con un concreto ánimo discriminatorio, o si una decisión supone una «diferencia de trato carente de fundamento objetivo y razonable» (STC 227/1998, de 26 de noviembre, FJ 4), o de si estamos ante un «tipo especial de desigualdad caracterizado por la naturaleza generalmente odiosa del prejuicio social descalificatorio»².

Estas tradicionales exigencias se tornan un poco más complejas, pues de lo que se trata es de evaluar de una forma minuciosa si, a lo largo del proceso de desarrollo e implantación del modelo, se han producido determinadas irregularidades, para descubrir qué herramientas o técnicas nos permiten subsanarlas.

Se ofrecerán, a continuación, algunas pinceladas de cómo el proceso de evaluación y examen debe cambiar si queremos tomarnos en serio la importante problemática de las discriminaciones, con especial referencia a aquellas que se producen en el campo del sistema penal, donde los bienes e intereses en juego son los más elementales, viéndose estos en entredicho ante determinadas decisiones como las propias de las herramientas de predicción del riesgo de violencia, cuyo ejemplo paradigmático (que nos servirá de base para explicar los genéricos problemas) es el de la conocida herramienta de predicción del riesgo de reincidencia COMPAS³. El titular que hizo famoso este algoritmo de predicción del riesgo de reincidencia formaba parte de una investigación publicada por la compañía ProPublica, y rezaba: «Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks»⁴.

Tras este titular, las alarmas de todos los críticos saltaron. El riesgo de que la tecnología acaparase un ámbito tan sensible como el criminal ya era motivo de preocupación. Pero las peores sospechas se habían confirmado; una máquina estaba tomando decisiones sobre la posibilidad de que un sospechoso fuera encarcelado en función del análisis de una serie de factores de riesgo, no demasiado conocidos. Este estudio pone de manifiesto que, además, las personas negras iban a tener más posibilidad de ser erróneamente clasificadas que las blancas. La polémica está servida. Pero ¿de dónde provienen realmente los problemas?

² RUIZ MIGUEL, A., (1996), «La igualdad en la jurisprudencia del Tribunal Constitucional», *Doxa. Cuadernos de Filosofía del Derecho*, 19, p. 53.

³ COMPAS es el acrónimo de la herramienta de predicción de la peligrosidad Correctional Offender Management Profiling for Alternative Sanctions.

⁴ ARGWIN, J. *et al.* (2016), «Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks», *Pro Publica*, 23 de mayo de 2016.

Las herramientas de tipo predictivo basadas en IA se nutren, como ya se ha indicado, de datos. Así, con un objetivo previamente marcado (en este caso, clasificar a las personas sospechosas de haber cometido un delito en función de su mayor o menor propensión a reincidir), se recogen unos datos sobre los que trabajar, se filtran, y se empieza a trabajar con algunos de los factores o atributos que se considera que mejor pueden alcanzar el objetivo de predicción marcado. Con estos datos y este objetivo prefijado, se comienza a entrenar el modelo y, una vez que se alcanzan unos parámetros de corrección o validez determinados, el modelo se testa sobre los datos.

A lo largo de este proceso pueden darse ya determinadas problemáticas que, aunque puedan parecer meros errores de cálculo, nos llevan ya a resultados intolerables. Veamos muy brevemente qué tipo de errores pueden aparecer en qué parte del proceso.

1.2.1. *Proceso de recolección y selección de datos*

Ya podemos encontrar algunos de los problemas más comunes; quizá el más conocido, el llamado «sesgo de representación», que se da cuando los datos no incluyen una muestra lo suficientemente representativa de la población que se está examinando. Por ejemplo, en los casos de reconocimiento facial, sería un ejemplo de este error el no seleccionar caras de personas negras o asiáticas para entrenar la base de datos. El modelo aprenderá sobre los datos que se le muestren; si nunca ha visto unos ojos rasgados, no los clasificará como ojos. Algo parecido ocurre con el caso de las mujeres cuyos currículos no son seleccionados. No es que las mujeres sean excluidas expresamente de los datos tomados como base del entrenamiento (ni siquiera el género era una variable tenida en cuenta como atributo), sino que, como consecuencia de un «sesgo histórico», los casos de personas que mayoritariamente llegan al éxito laboral son hombres, por lo que el modelo entiende que ese debe ser un atributo esencial para el éxito, y lo añade, indirectamente, entre los factores relevantes.

Ocurren también los conocidos como errores de medida, donde lo que sucede es una discrepancia entre la realidad que se pretende observar y la forma en que se representa en los datos. Así: el sesgo en los atributos (*feature bias*), cuando hay discrepancia entre los datos que se pretenden recoger como indicadores o atributos y los que realmente están disponibles (ejemplo: intentar conocer la tasa de delincuencia solo mediante las denuncias de un determinado delito, igualando el atributo delito cometido con el rasgo observable delito denunciado). El sesgo en la etiqueta (*label bias*), cuando hay discrepancia entre los datos y los resultados que se usan como medidor de corrección (ejemplo: como no podemos medir un dato de forma perfecta acudimos a uno que sea

aproximado, un *proxy*; ejemplo: para medir el indicador «éxito laboral» de una manera perfecta, habremos de seleccionar el proxy «haber sido objeto de ascensos»).

1.2.2. *Proceso de diseño del modelo*

Encontramos también algunos problemas o sesgos. Uno de los más típicos en los contextos relacionados con la justicia penal (o contextos sociales complejos en general) tiene que ver con el llamado *framing trap* o «sesgo de implementación». Ocurre cuando se produce una desconexión entre el problema que el modelo debe resolver y el entorno en que se usa (caracterizado por ser una estructura social compleja). Cuando el modelo interactúa en un entorno donde, además de los inputs y outputs clásicos, se pueden producir variaciones de estos factores como consecuencia del sistema social, los parámetros de justicia previamente establecidos pueden verse disminuidos o directamente eliminados. Por ejemplo, si una herramienta de predicción de la peligrosidad no tiene en cuenta que los jueces que van a aplicarla tienen sus propios sesgos, y en muchas ocasiones se desviarán de la decisión recomendada por el modelo por motivos diversos, dicha herramienta no será capaz de aportar una solución «justa» adaptada al contexto social real en que se va a implementar. Aparece en esta fase también un sesgo de tipo técnico, el denominado sesgo de aprendizaje (*learning bias*), que supone que ya solo elegir un modelo matemático sobre otro implica que se priorice un valor y no otro (por ejemplo, que se priorice la precisión global, en detrimento de la equidad, es decir, dando lugar a mayor disparidad en las decisiones).

1.2.3. *Momento de evaluación del modelo*

Deben examinarse en esta fase cuestiones de importancia mayúscula a los fines que aquí nos interesan. Una de las más complejas, sobre la que merece la pena un extenso debate que no puede desarrollarse aquí, será simplemente apuntada. Es necesario que, de cara a evaluar un modelo, se establezca de antemano un concepto de «justicia» o «equidad» prefijado, con el objetivo de poder evaluar si el modelo tiene un buen rendimiento. No todos son compatibles y no todos funcionan igual en cualquier contexto.

Así, manejaremos dos de los más usados a título de ejemplo (pues son, además, los dos conceptos de «equidad» en juego en el polémico supuesto COMPAS).

1.2.3.1. Paridad estadística o equidad de grupo

El modelo será justo, y no causará resultados discriminatorios intolerables, cuando las probabilidades de obtener un resultado positivo (ser clasificado como «peligroso») sean iguales para los dos grupos. Esto exige, por supuesto, prefijar qué grupos tienen aquellos atributos sensibles o que queremos proteger con mayor esfuerzo (resultan orientativas, quizá, las circunstancias que nuestra Constitución enuncia expresamente en su art. 14 CE como emanación del mandato de no discriminación). La raza es, sin duda, en este campo, uno de los atributos protegidos más incuestionables. Así, la pregunta a realizar sería: ¿es igual de probable para una persona blanca que para una negra ser clasificada dentro del umbral de peligrosidad? Así ocurría en el caso COMPAS. La proporción de acusados reincidentes era aproximadamente la misma con independencia de la raza.

1.2.3.2. Equilibrio de la tasa de error

Implica que exista igualdad de probabilidades de ser incorrectamente clasificado. Aquí es donde estaba el error en el famoso *leading case* COMPAS, pues en los casos en que la herramienta no predecía correctamente, el error afectaba más a un grupo que a otro; es decir, los acusados negros que finalmente no reincidían (falsos positivos) eran clasificados como más peligrosos que los acusados blancos que no lo hacían.

En ambos casos, lo que exige el modelo es tomar una decisión previa. Y ello nos sitúa en una posición donde queda un hueco (muy relevante) para la decisión humana, para la ética y, sobre todo, para el dilema. En definitiva, ¿qué errores nos importan más, o hacen más daño al sistema, en el concreto contexto de aplicación del modelo?, ¿aquellos que impliquen una clasificación incorrecta de la persona como peligrosa, con la consecuente privación de libertad de una persona injustamente, o aquellos que la clasifiquen incorrectamente como no peligrosa, provocando la liberación de una persona que acabará delinquir y causando un daño social? Este es realmente el tipo de dilema al que nos enfrentan las modernas herramientas tecnológicas.

Como todos los dilemas éticos, dichos retos requieren de las personas implicadas en el proceso de creación, implementación y evaluación una conciencia completa de la dificultad que los problemas de igualdad y discriminación suponen en contextos tan complejos. Lejos queda el supuesto tradicional en que el juez evalúa, con criterios meramente valorativos y abiertos, si una decisión puede ser discriminatoria por ser o no arbitraria, objetiva o razonable. La disciplina de la justicia algo-

rítmica requiere esfuerzos mucho mayores; requiere de la comunidad científica un interés por entender la complejidad de las implicaciones sociales de sistemas tan complejos como el penal, donde juegan principios y derechos esenciales, pero también un esfuerzo por parte de los juristas, que, indudablemente, deben trabajar en colaboración. Un problema complejo donde hemos de tener en cuenta las implicaciones éticas que tiene asumir un modelo u otro. Citando a Walzer, las diferentes «esferas de justicia» deben hacer que tomemos decisiones técnicas diferentes en función del contexto⁵.

2. ¿QUÉ INSTRUMENTOS NORMATIVOS SE HAN ENCARGADO DE LA DISCRIMINACIÓN ALGORÍTMICA?

Podemos distinguir dos grupos de instrumentos:

2.1. Instrumentos preventivos

Aunque podríamos nombrar muchos otros instrumentos normativos (ej.: la normativa sobre Protección de datos, la normativa sobre el mercado de servicios digitales o incluso alguna normativa sectorial que introduce pequeños avances, como la Ley rider⁶, o el Convenio [Marco] Internacional sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho del Comité sobre Inteligencia Artificial del Consejo de Europa, presentado en enero de 2023 y firmado por la Comisión Europea en septiembre de 2024), será prioritario centrarnos en el instrumento más importante por su aplicación no sectorial, sino general, y de carácter internacional. Nos referimos, en efecto, al Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial.

⁵ WALZER, M., (2016), *Las esferas de la justicia. Una defensa del pluralismo y la igualdad*. Fondo de Cultura Económica de España.

⁶ Nos referimos a la Ley 12/2021, de 28 de septiembre, por la que se modifica el texto refundido de la Ley del Estatuto de los Trabajadores, aprobado por el Real Decreto Legislativo 2/2015, de 23 de octubre, para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales. Su art. 64.4 del ET reconoce como derecho del trabajador a «ser informado por la empresa de los parámetros, reglas e instrucciones en los que se basan los algoritmos o sistemas de inteligencia artificial que afectan a la toma de decisiones que pueden incidir en las condiciones de trabajo, el acceso y mantenimiento del empleo, incluida la elaboración de perfiles», RIVAS VELASCO, M. J. (2023), «Uso ético de inteligencia artificial en justicia», *Diario La Ley*, 13. También: SENENT VIDAL, M. J. y VENTURA FRANCH, A. (2024), «Inteligencia artificial e igualdad de mujeres y hombres: una regulación jurídica necesaria. Especial referencia al proyecto de reglamento europeo», *IgualdadES*, 10, pp. 45-71.

2.1.1. *Reglamento de IA*

En su considerando 27 este instrumento normativo acoge explícitamente los principios éticos reconocidos por el HLEG⁷, entre los cuales se citan la diversidad, la no discriminación y la equidad. Se entiende que los sistemas de IA se desarrollan y utilizan de un modo que incluya a diversos agentes y promueva la igualdad de acceso, la igualdad de género y la diversidad cultural, al tiempo que se evitan los efectos discriminatorios y los sesgos injustos prohibidos por el derecho nacional o de la Unión.

Las diversas alusiones a la no discriminación sirven para hacer constar que existen determinados usos de las herramientas de predicción basadas en IA que suponen un «alto riesgo» y que, por tanto, deben estar sometidos a unos controles y requisitos mucho más exigentes. Aunque el listado específico se desarrolla en el anexo III, se ha venido llamando la atención, por considerarse especialmente sensibles, sobre los sistemas de IA utilizados para determinar el acceso a la educación, los usados para la gestión de trabajadores y acceso al empleo, o el acceso a determinados servicios y ayudas de carácter público o privado (poniendo como ejemplo el acceso a créditos), así como la gestión de migración y control fronterizo, y Administración de Justicia y procesos democráticos (todos ellos campos donde hemos podido ver ejemplos de sesgos y discriminaciones). En todo caso, la Comisión puede ampliar este listado.

Las obligaciones unidas a estos sistemas persiguen garantizar que el sistema de IA de alto riesgo funcione del modo previsto, en condiciones de seguridad, y que no se convierta en la fuente de alguno de los tipos de discriminación prohibidos por el derecho de la Unión. Se insta en el art. 9 lo que debe entenderse por sistema de gestión de riesgos (un proceso iterativo continuo, planificado y ejecutado a lo largo de todo el ciclo de vida de un sistema de IA de alto riesgo, que requiere una revisión y actualización periódicas y sistemáticas), mientras en el art. 10 se precisan las obligaciones en materia de gobernanza de datos.

Es en este art. 10 donde podemos ver reflejadas las exigencias derivadas del proceso centrado en el dato anteriormente explicado. Se señala la importancia del proceso de recogida y origen de los datos, la finalidad original, las operaciones de preparación (etiquetado, limpieza...), la formulación de hipótesis, la evaluación de la disponibilidad e idoneidad de los datos, el examen de posibles sesgos que puedan afectar a la salud y seguridad, repercutir negativamente en derechos fundamentales o dar lugar a discriminaciones prohibidas por el derecho de la Unión,

⁷ Acrónimo usado para referirse al High Level Expert Group designado por la Comisión Europea.

especialmente cuando los resultados de los daños influyan en las futuras operaciones. Se deben prever medidas adecuadas para detectar, prevenir y mitigar posibles sesgos e identificar lagunas en los datos.

Asimismo, se destaca la necesidad de que los datos sean variados y atiendan al contexto en que se van a aplicar: «Los conjuntos de datos tendrán en cuenta [...] las características o los elementos propios del entorno geográfico, contextual, conductual o funcional específico en el que está previsto utilizar el sistema de IA de alto riesgo». Y un detalle importante: se permite el uso, para precisamente detectar y mitigar sesgos, de categorías de datos especiales, poniéndose fin así a las posibles dudas sobre la pertinencia de usar algunos datos que por su contenido altamente personal y protegido pudieran quedar fuera en estricta aplicación de la normativa genérica de protección de datos. Esto ya se contemplaba en las versiones anteriores del Reglamento, pero con la definitiva se añaden algunos apartados posteriores al apartado 5 que no estaban antes y que van encaminados a limitar el tratamiento o almacenamiento o el uso de estos datos a los casos imprescindibles y justificados.

Parece que el art. 10 da con la clave para abordar los problemas de esta nueva forma de discriminación. Los artículos siguientes abordan otros aspectos que también pueden tener algo de relevancia para paliar el efecto discriminatorio, aunque quizá de forma más tangencial (exigencia de documentación técnica, obligaciones de información a usuarios, supervisión humana, solidez y ciberseguridad). Habremos de estar muy atentos, por tanto, a la aplicación cuidadosa del art. 10.

2.2. Instrumentos reactivos

2.2.1. Legislación antidiscriminación

Aunque podemos afirmar que es una mezcla entre preventivo y reactivo, la legislación antidiscriminación (desarrollada principalmente en las directivas europeas traspuestas a nuestro ordenamiento y en la Ley Integral 15/2022, de 12 de julio, para la igualdad de trato y la no discriminación, que tiene como ámbito objetivo de aplicación la IA y gestión masiva de datos), ofrece un marco del que partir. Dos son los conceptos principales que merece la pena analizar en la búsqueda de un marco de protección adecuado para los modernos ataques a la igualdad protagonizados por las herramientas tecnológicas: discriminación directa y discriminación indirecta.

Cabe matizar que, aunque la ley prevé la realización de un estudio de impacto sobre la igualdad y sobre el sesgo discriminatorio (art. 23.1), se trata en realidad más de una declaración de principios e intenciones, sin establecer el procedimiento para llevarla a cabo ni de quién depen-

de su ejecución, por lo que muy probablemente queda pendiente de un desarrollo reglamentario posterior. Su art.23 resulta especialmente novedoso, ya que establece obligaciones para las Administraciones públicas en relación con los algoritmos de IA para la toma de decisiones.

Sí que se aprecian, en cambio, avances al regular la Ley en su título iv las infracciones y sanciones en materia de igualdad de trato y no discriminación en general⁸. Su graduación va en relación con la gravedad de las acciones tipificadas, que incluyen desde la inobservancia de lo establecido en la Ley hasta los actos que constituyan discriminaciones múltiples. Se prevén multas que oscilan entre los 300 y los 500 000 euros.

Sin embargo, los conceptos centrales de discriminación sobre los que pivota la norma parecen ser insuficientes por las razones que a continuación se expondrán:

2.2.1.1. La inaplicabilidad de la discriminación directa

Este tipo de discriminación exige un trato menos favorable a un individuo por razón de su pertenencia a un colectivo protegido. Se pueden observar dos problemas:

— Esta definición no se corresponde con la lógica algorítmica, donde no es común que el resultado discriminatorio derive de que el programador use explícitamente la pertenencia al colectivo como input del modelo para asignarle una puntuación más baja. Los resultados discriminatorios quedan alejados de esta dinámica tradicional y se acercan más a los supuestos de sesgos sistemáticos arriba explicados.

— Exigir un trato desfavorable «por razón de la pertenencia a un colectivo protegido» desconoce que la dinámica algorítmica no se basa en la pertenencia de un individuo a un grupo, sino en la afinidad o correlación entre factores. Existen aquí dos noticias positivas: la jurisprudencia lleva tiempo reconociendo la discriminación por «afinidad», que no exige la pertenencia, y nuestra nueva Ley la reconoce expresamente.

⁸ ABA CATOIRA, A. (2023), «Discriminación en la comunicación y en la publicidad: comentario del art. 22 de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación». *IgualdadES*, 9, pp. 107-141. También: RODRÍGUEZ ÁLVAREZ, A. (2023), «La tutela procesal en la Ley integral para la igualdad de trato y la no discriminación». *IgualdadES*, 9, 175-204.

2.2.1.2. La fácil justificación de la discriminación indirecta

Pudiera parecer que la figura de la discriminación indirecta, que exige un trato «aparentemente neutro» con el que se ocasiona una desventaja particular, pudiera encajar algo más con la lógica algorítmica. Sin embargo, se presenta aquí otro inconveniente. Tradicionalmente, este tipo de discriminación está justificado siempre y cuando haya una finalidad legítima y los medios sean adecuados y necesarios. Es decir, se introduce el clásico test de proporcionalidad, que, en este caso, dificulta enormemente su aplicación en la dinámica propia de los algoritmos. Diversos autores, como Xenidis o Hacker⁹, apuntan que será fácil encontrar una finalidad legítima aludiendo a la mayor capacidad predictiva del algoritmo, así como también será fácil que el medio sea adecuado en cuanto prediga adecuadamente el fenómeno, y finalmente, será proporcional en sentido estricto en tanto no exista un medio menos gravoso, es decir, no se pueda acudir a una base de datos menos sesgada o el proceso de decisión algorítmico reduzca los sesgos igual o mejor, por poco que sea, que el proceso sin asistencia algorítmica.

En definitiva, se advierte por la doctrina que la legislación antidiscriminación es anticuada aún, pues no es capaz de recoger el fenómeno de las desprotecciones causadas por el sesgo algorítmico. Tendremos que esperar a comprobar si esta novedosa ley, donde ya aparecen términos que la crítica venía exigiendo, tales como discriminación por asociación o interseccional, es capaz de arrojar algo de luz sobre el fenómeno. Finalmente, se ha criticado por algunos autores que solo se aplique a la Administración pública, quedando fuera los entes privados, aunque sí que dice que se promoverá la creación de un sello de calidad en los algoritmos¹⁰.

2.2.2. *Proposición de Ley Orgánica de regulación de las simulaciones de imágenes y voces de personas generadas por medio de la inteligencia artificial. Presentada por el Grupo Parlamentario Plurinacional SUMAR*

El texto aborda la necesidad de proteger los derechos al honor, la intimidad y la propia imagen ante el uso de inteligencia artificial

⁹ XENIDIS, R. (2020), «Tuning EU equality law to algorithmic discrimination: Three pathways to resilience», *Maastricht Journal of European and Comparative Law*, 27(6), pp. 736-758. También, HACKER, P. (2018), «Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law», *Common Market Law Review*, 55(4), pp. 1143-1186.

¹⁰ FLORES ANARTE, L. (2023), «Sesgos de género en la inteligencia artificial: El Estado de derecho frente a la discriminación algorítmica por razón de sexo». *Revista Internacional de Pensamiento Político*, 18(18), pp. 95-120.

para recrear imágenes y voces sin consentimiento. Se señala que el Parlamento Europeo ya ha destacado los riesgos de los *deepfakes*, como la difamación, el fraude y el acoso, señalando que la mayoría de estos contenidos son de carácter pornográfico, afectando mayormente a mujeres y constituyendo una forma de violencia sexual.

El texto también menciona la Propuesta de Reglamento de IA, que sugiere regular la transparencia de sistemas de IA de alto riesgo. Además, se plantea la modificación del Código Penal español para penalizar la creación y difusión de imágenes o audios generados por IA que vulneren el honor o la dignidad de las personas, considerando estos actos como injurias públicas.

Se introducen medidas para la retirada rápida de estos contenidos, así como cambios en la Ley de Protección de Datos, la Ley de Competencia Desleal y el Régimen Electoral, sancionando la difusión de *deepfakes* durante procesos electorales. Además, se crean consejos para supervisar el uso de IA.

Sin embargo, la doctrina ha reaccionado de forma crítica a la publicación de esta propuesta, señalando algunos de sus errores más patentes:

En primer lugar, se afirma que es demasiado precipitada a la hora de regular un contenido que aún no se ha llegado a aplicar de forma uniforme en la UE (quizá sería mejor esperar a que el Reglamento comience a aplicarse). De hecho, el propio Reglamento no califica las tecnologías *deepfake* como de alto riesgo, sino dentro de la categoría de riesgo sistémico, por lo que las precauciones sobre gobernanza de datos, que son las más dirigidas a paliar los vicios de igualdad de las herramientas, no serían de aplicación obligatoria¹¹.

También se ha señalado que, pese a resaltarse en la exposición de motivos de la Propuesta que las violencias del *deepfake* son sufridas principalmente por mujeres en contextos que hacen ver que son en realidad una clase de violencia sexual, no se tratan como tal, sino que se regulan de forma genérica como delitos contra el honor, sin hacer referencia a este posible desvalor añadido¹².

¹¹ SANTISTEBAN GALARZA, M., (2024), «La criminalización de las ultrafalsificaciones (con especial atención a las implicaciones de la normativa europea de servicios digitales e inteligencia artificial)», *Revista de Derecho Penal y Criminología*, 3.ª época, 31, pp. 211-249.

¹² *Idem*. Debe señalarse que en las propuestas legislativas más recientes en España se aprecia una clara tendencia a configurar los *deepfakes* como una forma de vulneración del derecho al honor. Así, el Anteproyecto de Ley Orgánica de protección civil del derecho al honor, a la intimidad personal y familiar y a la propia imagen, aprobado en primera vuelta por el Consejo de Ministros el 13 de enero de 2026, introduce expresamente la utilización de imágenes o voces generadas mediante IA sin consentimiento como una intromisión ilegítima en estos derechos fundamentales, incluyendo por primera vez los *deepfakes* en el catálogo de conductas lesivas del honor. En paralelo, el Proyecto de Ley Orgánica para la protección de las personas menores de edad en los entornos digitales (iniciativa presentada el 27 de marzo de 2025 y actualmente en

Finalmente, se destaca que el texto no concuerda con lo que señala la recientemente aprobada (en vigor desde el 13 de junio de 2024) Directiva sobre la lucha contra la violencia contra las mujeres y la violencia doméstica. Este nuevo instrumento, pendiente de trasposición en España, aborda la ciberviolencia no cubierta por el Convenio de Estambul. Establece normas mínimas para delitos como la difusión no consentida de material íntimo o manipulado, el ciberacecho, el ciberacoso y la incitación al odio online. Su art. 7 pretende la criminalización de la difusión de contenido íntimo o manipulado (*deepfakes*), siguiendo un modelo similar al del Reino Unido, proponiendo castigar a quienes creen o manipulen imágenes sexuales sin consentimiento, y amplía el concepto de «difusión», eliminando el requisito de que el contenido llegue a una multitud de personas. Además, penaliza a cómplices e inductores de estos delitos, obliga a castigar la tentativa y obliga a establecer penas que impliquen, como mínimo, un año de prisión. Precisamente en este último punto, relativo a la pena mínima a imponer por este delito, la propuesta de ley orgánica presentada en España no es acorde con los mínimos exigidos por la UE.

Será necesario revisar detenidamente los términos de una propuesta como la presente si es que finalmente se aborda la tipificación del *deepfake* a nivel nacional.

2.2.3. *Delitos diseminados a lo largo del CP*

Tenemos en el Código Penal algunos ejemplos de tipos delictivos que castigan discriminaciones, principalmente divididos en dos grandes grupos: los delitos contra la Constitución y los delitos contra los trabajadores.

Así, podemos distinguir entre los delitos de denegación de una prestación por motivos discriminatorios de los artículos 511 y 512 CP¹³ dentro de los delitos contra la Constitución. Autores como Valls¹⁴

tramitación parlamentaria) refuerza esta misma lógica al prever la sanción de las denominadas “ultrafalsificaciones” o *deepfakes*, especialmente cuando afectan a menores o tienen contenido sexual. En conjunto, ambos textos evidencian una progresiva consolidación doctrinal y normativa de los *deepfakes* como ataques al honor —y no solo como ilícitos tecnológicos o penales—, aunque se trata todavía de una cuestión abierta y sometida a debate legislativo y doctrinal, dado que ambas iniciativas se encuentran en fase prelegislativa o de tramitación y requieren aún su aprobación definitiva por las Cortes Generales.

¹³ GARROCHO SALCEDO, A. (2013), «Delitos de denegación de una prestación por motivos discriminatorios de los artículos 511 y 512 CP». *Eunomía. Revista en Cultura de la legalidad*, 5.

¹⁴ VALLS PRIETO, J. (2023), «La inteligencia artificial frente a la colectividad. Una protección de los bienes jurídicos colectivos frente al auge de los sistemas inteligentes». *Estudios penales y criminológicos*, 44(Ext.), p. 20 y ss. Aunque acepta las dificultades que puede implicar aplicar los delitos previstos desde el paradigma de la comisión dolosa.

han puesto de manifiesto la posibilidad de aplicar estos delitos en casos de discriminaciones algorítmicas realizadas por la Administración.

El segundo grupo se enmarca dentro de la protección de los trabajadores. Este último delito de discriminación en el empleo público (314 CP¹⁵), de escasa aplicación, que requiere además que se haya notificado algún tipo de requerimiento previo tras haberse constatado el hecho discriminatorio¹⁶, parece no ser tampoco la mejor forma de luchar contra la discriminación algorítmica.

3. CONCLUSIONES

Para abordar el fenómeno de la igualdad, el sesgo y la discriminación dentro de la lógica de la IA debemos conocer cómo funciona el novedoso proceso de creación de estos sistemas, y ser conscientes de que, quizá, la clave esté en adelantar el momento de actuación por parte de los poderes públicos. No se trata tanto de engordar la normativa reactiva, es decir, regular nuevas formas de castigo, especialmente penal, en un campo donde los elementos subjetivos normalmente requeridos son difíciles de afirmar y donde, hasta ahora, han encontrado difícil amparo los supuestos problemáticos, como de reaccionar en fases anteriores. Para ello es imprescindible conocer la dinámica del sesgo, cómo cada tipo de sesgo puede generarse en un momento específico condicionado por diferentes factores.

Una última idea, esencial, más allá del giro que parece que todas las normas preventivas están tomando en cuanto al establecimiento de deberes (más o menos obligatorios) de evaluación de impacto y gobernanza de datos en todo su ciclo de vida, nada de esto servirá si renunciamos a la transparencia y la explicabilidad.

Se puede vislumbrar a la perfección la importancia de este aspecto al analizar un reciente estudio encargado por una ONG de ámbito europeo e internacional, que pretende realizar un estudio comparativo respecto de los riesgos que las herramientas de policía predictiva utilizadas en distintos Estados europeos pueden plantear para los dere-

¹⁵ ABADÍA SELMA, A. (2023), «La protección penal de los trabajadores frente a la inteligencia artificial en el ámbito del delito de discriminación laboral», *Revista General de Derecho Penal*, 39.

¹⁶ *Idem*, explicando las dificultades interpretativas y aplicativas del precepto lo relaciona con la IA. Destaca en p. 33: que existen «procesos de selección de personal que están desde un inicio totalmente “viciados” por la inteligencia artificial y los algoritmos con sesgos, pudiendo producirse una discriminación incluso *ex ante*, sin que todavía el sujeto pasivo haya entrado a formar parte de la plantilla de trabajadores».

chos fundamentales de los ciudadanos¹⁷. La profesora Martínez Garay, una de las encargadas de conducir este estudio y mostrar sus resultados en el caso español, después de analizar nuestras herramientas de predicción más renombradas, señala cómo del estudio detallado y profundo de los datos es posible inferir que sistemas como Riscanvi, usado en el ámbito penitenciario catalán, no discriminan a las personas racializadas, pero sí lo hace respecto de las personas con trastornos mentales y adicción a sustancias. Solo fomentando el análisis exhaustivo y abierto de las herramientas y su funcionamiento podremos, primero, conocer cuál es la realidad, y solo a partir de este conocimiento podremos, en su caso, tomar medidas ulteriores. La lucha por la justicia y la igualdad debe empezar por la lucha por el conocimiento. Sin transparencia y explicabilidad difícilmente podremos hablar de igualdad.

4. BIBLIOGRAFÍA

- ABA CATOIRA, A. (2023), «Discriminación en la comunicación y en la publicidad: comentario del art. 22 de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación». *IgualdadES*, 9, 107-141. Accesible en: <https://doi.org/10.18042/cepc/IgdES.9.04>.
- ABADÍA SELMA, A. (2023), «La protección penal de los trabajadores frente a la inteligencia artificial en el ámbito del delito de discriminación laboral», *Revista General de Derecho Penal*, 39.
- ARGWIN, J., LARSON, J., MATTU, S., KIRCHNER, L. (2016), «Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks», *Pro Publica*, 13 mayo de 2016. Accesible en: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [06/02/2025].
- BRAVO BOLADO, A., (2023), «Justicia algorítmica. Un enfoque sociotécnico». *Revista de Estudios Penales y Criminológicos*, núm. 44.
- FLORES ANARTE, L. (2023), «Sesgos de género en la inteligencia artificial: El Estado de derecho frente a la discriminación algorítmica por razón de sexo». *Revista Internacional de Pensamiento Político*, 18(18), pp. 95-120.
- GARROCHO SALCEDO, A. (2013), Delitos de denegación de una prestación por motivos discriminatorios de los artículos 511 y 512 CP. *Eunomía. Revista en Cultura de la legalidad*, (5).
- HACKER, P. (2018), «Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law». *Common Market Law Review*, 55(4), pp.1143-1186.
- MARTÍNEZ GARAY, L., et al. (2024), «Three predictive policing approaches in Spain: VioGén, Riscanvi and Veripol. Assessment from a human rights perspective».

¹⁷ MARTÍNEZ GARAY L., et al. (2024), «Three predictive policing approaches in Spain: VioGén, Riscanvi and Veripol. Assessment from a human rights perspective».

- PRIETO, J. V. (2023), «La inteligencia artificial frente a la colectividad. Una protección de los bienes jurídicos colectivos frente al auge de los sistemas inteligentes». *Estudios penales y criminológicos*, 44(Ext.).
- RIVAS VELASCO, M. J. (2023), «Uso ético de inteligencia artificial en justicia. Diario La Ley, 13/07/2023». *Diario la Ley*, 23 de junio de 2023. Disponible en: <https://tinyurl.com/4537acrz>.
- RODRÍGUEZ ÁLVAREZ, A. (2023), «La tutela procesal en la Ley integral para la igualdad de trato y la no discriminación». *IgualdadES*, 9, 175-204. Accesible en: <https://doi.org/10.18042/cepc/IgdES.9.06>.
- RUIZ MIGUEL, A., (1996), «La igualdad en la jurisprudencia del Tribunal Constitucional». *Doxa. Cuadernos de Filosofía del Derecho*, (19), 39-86.
- SANTISTEBAN GALARZA, M. (2024), «La criminalización de las ultrafalsificaciones (con especial atención a las implicaciones de la normativa europea de servicios digitales e inteligencia artificial)», *Revista de Derecho Penal y Criminología*, 3.ª época, 31, pp.211-249.
- SEMENT VIDAL, M. J. y VENTURA FRANCH, A. (2024), «Inteligencia artificial e igualdad de mujeres y hombres: una regulación jurídica necesaria. Especial referencia al proyecto de reglamento europeo». *IgualdadES*, 10, pp.45-71.
- XENIDIS, R. (2020), «Tuning EU equality law to algorithmic discrimination: Three pathways to resilience» *Maastricht Journal of European and Comparative Law*, 27(6), pp.736-758.

Anexo de normativa y sentencias

- Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Directiva sobre la lucha contra la violencia contra las mujeres y la violencia doméstica.
- Ley 12/2021, de 28 de septiembre, por la que se modifica el texto refundido de la Ley del Estatuto de los Trabajadores, aprobado por el Real Decreto Legislativo 2/2015, de 23 de octubre, para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales.
- Ley Integral 15/2022, de 12 de julio, para la igualdad de trato y la no discriminación.
- Proposición de Ley Orgánica de regulación de las simulaciones de imágenes y voces de personas generadas por medio de la inteligencia artificial. Presentada por el Grupo Parlamentario Plurinacional SUMAR.
- Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) núm. 300/2008, (UE) núm. 167/2013, (UE) núm. 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial).
- STC 227/1998, de 26 de noviembre.

