SOBRE *FAIRNESS* Y *MACHINE LEARNING:* EL ALGORITMO ¿PUEDE (Y DEBE) SER JUSTO? *

On Fairness and Machine Learning: Can (and Should) the Algorithm Be Fair?

Nuria Belloso Martín **

Anales de la Cátedra Francisco Suárez
Fecha de recepción: 25/7/2022
ISSN: 0008-7750, núm. 57 (2023), 7-38
Fecha de aceptación: 22/08/2022
http://dx.doi.org/10.30827/ACFS.v57i.25250

RESUMEN El uso cada vez más frecuente de la Inteligencia Artificial en el ámbito del Derecho, obliga a plantearse si las decisiones automatizadas pueden, y deben, ser justas. El algoritmo, en el *Machine Learning,* tiene la virtualidad de ir aprendiendo, lo que lo dota de un cierto grado de autonomía. Sesgos, discriminaciones y desigualdades que derivan de decisiones automatizadas, ponen al descubierto el mito del algoritmo justo. El criterio de justicia que se exige en la concepción analógica del Derecho también debe exigirse en la dimensión digital. En este trabajo, desde la dificultad inicial de una falta de consenso sobre qué sea la *fairness*, examino cómo incorporar la *fairness* al algoritmo. Ello exigirá un previo análisis de los fundamentos iusfilosóficos y de algunas de las teorías de la justicia (utilitaristas, contractualistas, comunitaristas, igualitaristas) a partir de las cuales se puedan establecer parámetros, correctores y garantías para la consecución de la imprescindible correlación entre la *fairness* artificial y la *fairness* legal.

Palabras clave: *Fairness, Machine Learning,* Algoritmo, Sesgo, Igualdad.

ABSTRACT The increasingly frequent use of Artificial Intelligence in the field of law, forces us to consider whether automated decisions can, and should, be fair. The algorithm, in *Machine Learning,* has the potential to learn, which gives it a certain degree of autonomy. Biases, discriminations and inequalities that derive from automated decisions show the myth of the fair algorithm. The standard of justice that is required in the analogical conception of Law must also be required in the digital dimension. In this paper, from the initial difficulty of a lack of agreement on what *fairness* is, I examine how to incorporate *fairness* into the algorithm. This will require a previous analysis of the legal philosophical foundations and some of the theories of justice (utilitarians, contractualists, communitarians, egalitarians) from which parameters, correc-

---

* Para citar/citation: Belloso Martín, N. (2023). Sobre *fairness* y *marchine learning*: el algoritmo ¿puede (y debe) ser justo? *Anales de la Cátedra Francisco Suárez* 57, pp. 7-38.

** Universidad de Burgos. Departamento de Derecho Público. Hospital del Rey, s/n, 09001 Burgos (España). Correo electrónico: *nubello@ubu.es*

tors and guarantees can be established to achieve the essential correlation between artificial *fairness* and legal *fairness*.

Keywords: Fairness, Machine Learning, Algorithm, Bias, Equality.

> "Probablemente nunca tengamos una definición simple y universalmente aceptada de lo que hace que un algoritmo sea justo. Esta pregunta no es estrictamente técnica, es ética" (O'Neil).

1. Introducción

La Inteligencia Artificial (en adelante, IA), término acuñado por J. McCarthy en 1956, para nombrar lo que antes se llamaba "simulación computerizada", está siendo objeto de investigación preferente en esta última década, cuyos resultados han dado lugar a enriquecedores debates con respecto a la relación entre IA e inteligencia humana, sobre su aplicación y usos en áreas muy diversas, y sobre si sus efectos beneficiosos consiguen primar sobre las amenazas y males que también pueden desencadenar. La primera relación entre la IA y el Derecho fue llevada a cabo por Buchanan y Headrick (1970), cuando se plantearon de modo directo y concreto la cuestión de si el razonamiento jurídico era susceptible de ser computable. Desde entonces, se han realizado múltiples trabajos y análisis sobre esta especialidad.

En la publicación de Cathy o Neil, *Armas de destrucción matemática. Cómo el big data aumenta la desigualdad y amenaza la democracia* (2018) ya se cuestionaba el rol que los sistemas de decisión automatizados representaban en el funcionamiento diario de la sociedad (compañías aseguradoras, concesiones crediticias, departamentos de empleo/contratación, Administración pública y políticas públicas, prevención y control del fraude, tratamiento de datos sanitarios) (Eubanks, 2021) que acaban repercutiendo en la interpretación del riesgo y en la predicción-evaluación de daños a derechos fundamentales[1]. Con la utilización generalizada de modelos de aprendizaje automático en nuestra vida diaria, los investigadores han reflexionado sobre cómo equilibrar las ventajas que aportan con respecto a sus también efectos negativos. Los usos de la IA en el Derecho suponen un

1. El nivel de riesgo es uno de los pilares sobre los que la Unión Europea viene trabajando sobre la IA, como se pone de manifiesto en la Propuesta de Reglamento Europeo sobre el uso de la Inteligencia Artificial, de abril de 2021.

reto para la ética y para el Derecho antidiscriminatorio en particular, retos que acaban confluyendo en la necesidad de que la justicia presida el diseño, el funcionamiento y las decisiones que se adoptan a partir del algoritmo. Reducir o minimizar sesgos, velar por el respeto al principio de igualdad de oportunidades, neutralizar prejuicios, resolver la dicotomía discriminación directa/discriminación indirecta, son sólo una muestra de las aplicaciones concretas de la exigencia de justicia algorítmica y de la búsqueda de equidad e imparcialidad en la toma de decisiones.

Para evitar malentendidos conceptuales, hago la advertencia preliminar de que justicia y equidad no son términos sinónimos (Goldman y Cropanzano, 2015), al igual que tampoco lo son, a su vez, equidad e imparcialidad. Precisar tales conceptos requeriría un estudio profundo que ahora excedería los límites propuestos. A riesgo de simplificar en exceso, considero que la justicia, en su conceptualización de valor y fin del Derecho, sería la categoría omnicomprensiva tanto de la equidad como de la imparcialidad, entendiendo por estos criterios instrumentales el enfoque dirigido a lograr el fin más amplio que es la justicia. Puesto que lo que me propongo en este trabajo es justificar que es necesaria tanto la justicia como la *fairness* (bien se entienda como equidad o como imparcialidad) en el algoritmo, y reflexionar sobre los modelos que se han propuesto para reducir la "injusticia" con la máxima correlación basada en la optimización de la equidad, en este trabajo los utilizaré indistintamente. No hay una definición clara ni consenso unívoco sobre *fairness* porque de las diversas propuestas que existen, encuentran su sentido según el contexto al que se apliquen, es decir, son aplicables parcialmente a contextos distintos. Aquí reside una de las principales razones de la dificultad de precisar qué sea *fairness.*

La *fairness* algorítmica es actualmente un subcampo del *Machine Learning* (ML) en rápido desarrollo, y destacados investigadores, principalmente del ámbito anglosajón, están trabajando en este ámbito. Con todo, a pesar del creciente número de publicaciones e investigaciones en curso, todavía hay una falta de literatura crítica que explique la interacción del ML con las ciencias sociales de la filosofía, la sociología y el Derecho. En unos casos, hay nociones asociativas de *fairness* que son definidas en términos de correlación o dependencia entre variables, por ejemplo, paridad demográfica (Dwork *et al.,* 2012), probabilidades igualadas (Hardt *et al.,* 2016) y paridad predictiva (Chouldechova, 2016; Zafar *et al.,* 2017); en términos del ámbito de aplicación, se trabaja tanto sobre la *fairness* a nivel de grupo, como por ejemplo, probabilidades igualadas (Hardt *et al.,* 2016) e igualdad de esfuerzo (Huang *et al.,* 2020), como también sobre nociones de *fairness* a nivel individual, como equidad individual (Dwork *et al.,* 2012) o equidad contrafáctica (Kusner *et al.,* 2017). En cuanto a las técnicas para eliminar

o suprimir discriminación, existen enfoques de preprocesamiento (Zhao *et al.,* 2020; Zhao, 2021), enfoques de procesamiento (Zafar *et al.,* 2017) y enfoques de posprocesamiento (Hardt *et al.,* 2016; Dwork *et al.,* 2018). Hay también explicaciones sobre las opciones disponibles para cuantificar la discriminación y hacer cumplir la equidad en trabajos que muestran encuestas recientes (Verma y Rubin, 2018; Caton y Haas, 2020; Mehrabi *et al.,* 2021), así como una investigación sobre actitudes públicas hacia diferentes nociones (Saxena *et al.,* 2019).

Sin embargo, la filosofía y los contenidos metodológicos de las consideraciones de equidad subyacentes a menudo no están claramente articulados. De ahí que no resulte sorprendente ver que las nociones de *fairness* propuestas por la comunidad de aprendizaje automático se hacen eco de ciertas consideraciones de justicia y han dirigido su mirada a filósofos morales y políticos. Varios trabajos recientes han señalado la necesidad de reflexionar sobre tales conexiones (Binns, 2018; Barocas *et al.,* 2020).

La IA es un conjunto de innovaciones tecnológicas, cuyo funcionamiento debe desarrollarse ajustándose a unos parámetros ético-jurídicos (Pérez Luño, 1996; Llano Alonso, 2018; Solar Cayón, 2020; Campione, 2020; De Asís, 2022; Martínez García, 2020; Añón Roig, 2022, entre otros). La doctrina ha señalado cinco macro-principios fundamentales (Floridi, 2022), siendo cuatro comunes a la bioética: *beneficiencia, no maleficencia, autonomía* y justicia. A estos se añade un quinto principio, el de la explicabilidad, es decir, la IA debe ser inteligible y responsable (*accountability*). Sobre uno de estos cinco principios, el de la justicia, aplicada al aprendizaje automático, en su vertiente de imparcialidad, es sobre el que me propongo ofrecer unas reflexiones, para lo que analizaré distintas cuestiones que permitan profundizar en la pretendida equidad (*fairness*) algorítmica.

Las dificultades de qué sea la justicia y qué sea lo justo, que han inquietado a los juristas durante siglos en el ámbito del Derecho, se trasladan ahora a todas las áreas en las que se aplica la Inteligencia Artificial. Una primera pregunta que surge es la de bajo qué concepción de la justicia se debería de diseñar el algoritmo: ¿con qué teoría de la justicia habría que programarlo? ¿Con las ideas de Aristóteles, de Bentham, de Kant, de Marx, de Kelsen, de Sandel, de MacIntyre o de Rawls? ¿Qué perspectiva debería de tener: utilitarista o personalista? ¿Individual-liberalista o comunitarista? ¿Y cómo debería hacerse para cargar en el sistema el concepto tan esquivo —pero a la vez necesario— de "justicia social"? ¿Cómo se actualizaría el algoritmo en la materia? ¿Quién sería el llamado a hacerlo? (Krenz, 2021,7). Justicia climática, justicia distributiva, justicia retributiva, justicia global, justicia transicional, justicia intergeneracional (Miller, 2021) y, ahora, ¿justicia algorítmica? En realidad, esta última no constituye un tipo de justicia en

sí misma sino un cauce diferente (a través de la IA) al que se podrá recurrir para lograr la consecución de todas las anteriores. Las teorías de la justicia permitirán reflexionar sobre la literatura actual que versa sobre *fairness* algorítmica (Miller 2021).

Un aspecto a dilucidar es el de si hay identidad o existen divergencias entre el concepto de la *fairness* algorítmica y el de la *fairnes*s legal. Por ejemplo, un algoritmo equitativo, ¿podría potencialmente aplicar la forma de 'acción positiva' o 'acción afirmativa' con el propósito de compensar directamente alguna categoría en desventaja? La introducción de mecanismos correctores —como una variable correctora— equivale a exigir al Derecho respuestas a nuevas problemáticas y amenazas a los derechos humanos, pero el Derecho no está preparado para dar una respuesta híbrida.

A partir de la relación entre *fairness* e IA, en este estudio me propongo identificar el concepto de *fairness* algorítmica, entre las numerosas definiciones que se han formulado. Me limitaré a una de las modalidades de aprendizaje, el *Machine Learning* (aprendizaje automático) y analizaré si es posible la consecución de la aspiración a la equidad ('*fairness*'), para lo que tomaré en consideración las diversas métricas y criterios que se pueden utilizar para configurar una *fairness* algorítmica, poniendo de manifiesto la existencia de brechas entre las mediciones de *Machine Learning* y la compleja realidad del ser humano sociotécnico (Xiang, Raji, 2019). Entre las dificultades para lograr la imparcialidad en los procesos decisionales algorítmicos destacaré la problemática de los sesgos ('*bias*'). Identificar y mitigar los sesgos de la IA continúa siendo un reto, pero lo más urgente es reducir la probabilidad de resultados indeseables. Tratar a todos los individuos y colectivos de una manera justa —con la carga jurídica que conlleva en cuanto a la exigencia de igualdad de oportunidades o de discriminación positiva— pone de manifiesto que, al igual que hay tensiones en el mundo analógico, el problema se repite en el mundo digital. A partir de ahí, revisaré cuatro modelos de teorías de la justicia, que permiten entender la aplicabilidad (o no) de la *fairnees* digital.

2. Algunas precisiones sobre algoritmos y *Machine Learning* (para poder llegar a la definición de *fairness* algorítmica)

Comienzo realizando unas precisiones conceptuales sobre algoritmo, *fairness* y *Machine Learning*. Un algoritmo informático consiste en un conjunto de instrucciones definidas, secuenciadas, ordenadas y acotadas para resolver un problema, realizar un cálculo o desarrollar una tarea. Un programa informático es un conjunto de algoritmos ordenados y codificados

en un lenguaje de programación para poder ser ejecutado. Los algoritmos son precisos, sin ambigüedad, ordenados, finitos y concretos.

Machine Learning (ML) y *Deep Learning* (DL) son los dos conceptos clave sobre los que se sustenta la ciencia de datos y que se inscriben en el campo más amplio de la Inteligencia Artificial. El aprendizaje automático (ML) es una subcategoría de IA que se refiere al proceso por el cual los PC desarrollan el reconocimiento de patrones o la capacidad de aprender continuamente y realizar predicciones basadas en datos, tras lo cual realizan ajustes sin haber sido programados específicamente para ello. Explicado de forma breve, ML es "el área de la IA que desarrolla programas informáticos capaces de aprender por sí mismos y realizar predicciones" (Barona Villar, 2021, p. 97). Tales sistemas de aprendizaje automático hacen un uso cada vez más amplio de una gran cantidad de datos sobre el comportamiento humano que van recogiendo desde diversos canales (social media, app, datos telefónicos, transacciones a través de tarjetas de crédito, etc.). El aprendizaje automático utiliza algoritmos para analizar los datos, aprender de ellos y tomar decisiones informadas a partir de aquellos que han aprendido. Por su parte, el aprendizaje profundo estructura algoritmos en niveles para crear una "red neuronal artificial" capaz de aprender y tomar decisiones inteligentes por sí misma, pudiendo realizar tareas más complejas que el ML. Incido en que un algoritmo de ML puede procesar datos y desarrollar un autoaprendizaje de alguna manera autónoma, lo que desvirtúa o, al menos, exige matizar el argumento de que una IA se sujeta estrictamente a su programación. Las diversas cuestiones a las que voy a hacer referencia en este trabajo se limitan a la ML y con relación a su capacidad de toma de decisiones a partir de los patrones y modelos con que la IA ha sido entrenada[2].

---

2. Machine Learning se apoya en un proceso de enseñanza-aprendizaje, que consiste en suministrarle a la computadora grandes volúmenes de datos para que esta aprenda de forma automática a realizar una determinada tarea o función. Por ello, Machine Learning se relaciona con el análisis big data y el data science. A estas sesiones de enseñanza-aprendizaje se les conoce como entrenamientos. Así, en su aplicación al ámbito jurídico (*legaltech*) si se pretende que un algoritmo de machine learning aprenda a reconocer los expedientes jurídicos, entonces se realizan sesiones de entrenamiento en el que se le muestra a dicho algoritmo millones de documentos, entre expedientes y no expedientes, para que aprenda a diferenciarlos, clasificarlos, etiquetarlos y a trabajar con ellos. Las aplicaciones de ML al sector legal son numerosas: análisis predictivos, análisis causal, análisis de contenido, etc. En definitiva, cada vez son más sus aplicaciones al sector de la abogacía (Barrio Andrés, 2019) y al ámbito legal en general (Solar Cayón, 2019). El debate que se plantea en este trabajo afectaría a todos estos ámbitos susceptibles de aplicación de ML ya que toda disquisición legal estará influida por la *fairness*.

En su forma más básica, el aprendizaje automático utiliza algoritmos programados que reciben y analizan datos de entrada para predecir los valores de salida dentro de un rango aceptable[3]. La cantidad y calidad de los datos, así como otros parámetros (lo variado que sea el conjunto de datos de entrenamiento, la "suciedad" o "ruido" en los datos, repeticiones, sesgos, el hecho de que lleven información adicional —es decir que hayan experimentado un proceso de anotación de datos, etc.—) tendrán su influencia en el ML. La IA trabaja a partir del algoritmo, integrado por tres elementos constitutivos (*input*; procedimiento; *output*) el cual puede definirse, de forma elemental, como un procedimiento codificado para transformar el *input* (datos) en *output* (resultado esperado) mediante una serie de cálculos. En la primera fase, una de las principales preocupaciones es la presencia de errores en el *dataset* del *input*; en la segunda fase, un tema debatido es el de la transparencia y accesibilidad del procedimiento —el denominado *black box*—[4]; y en la tercera fase, el aspecto problemático es el de los posibles efectos discriminatorios de la decisión algorítmica. Un problema puede afectar a una o más de estas tres fases al mismo tiempo.

Por último, con respecto a la *fairness*, opto por recurrir al término en inglés, en la mayor parte de los casos, dado que no hay consenso en cuanto a su significado preciso, utilizándose como imparcialidad, como equidad y como justo. Sin embargo, *"justice"* es el vocablo que fielmente reflejaría la justicia. La idea que subyace es común a todos estos significados, pero los matices son diferentes. La mayor parte de los investigadores en la materia también utilizan "*fairness*". Imparcialidad sería uno de los términos más

3. El Big Data consiste en el almacenamiento y procesamiento de cantidades masivas de datos con gran potencial para ser extraídos y organizados de forma que proporcionen información valiosa para, en este caso, el ámbito jurídico. El Big Data actúa como un *input* que recibe un conjunto masivo de datos que necesitan ser procesados y los estandariza para convertirlos en útiles. A partir de ahí, se desencadena la virtualidad de la IA, que de forma simple se define como un conjunto de softwares que aprovechan el *output* generado por estos resultados para crear series de algoritmos que hacen que programas y mecanismos puedan mostrar comportamientos inteligentes y razonar como lo hacen los humanos. La IA aplicada al Big Data permite, además de reconocer patrones y probabilidades de resultado futuro, detectar desviaciones, identificando las anomalías que salgan del rango establecido. Por ejemplo, uno de esos rangos es el criterio de *fairness*.

4. Ya se ha explicado que la rama ML de la IA se basa en el aprendizaje de una tarea a partir de miles de datos. Entre las características de algunos de los algoritmos ML, entre los que están las redes neuronales, se encuentra el "black-box" (caja negra): el algoritmo hace buenas predicciones, pero no se sabe con detalle el procedimiento o *iter* que ha seguido para llegar, por ejemplo, a conceder una determinada ayuda social a una familia y no a otra, aparte de que haya aplicado los criterios legales que se contemplaran en la convocatoria de la ayuda.

aceptados, aunque como explicaré más adelante, es susceptible de matizaciones (imparcialidad individual, grupal, igual oportunidad y otras variantes).

2.1. El mito del algoritmo justo

Por regla general, hay una tendencia equivocada a creer que los procesos decisionales y las decisiones que emanan de una IA resultan neutrales o, como mínimo, más imparciales y equitativos que si derivaran de una decisión humana. Cada vez son más los investigadores que buscan soluciones para superar los problemas de la discriminación en los sistemas de software automatizados y que se preocupan por incorporar la idea de *fairness* en la estructura del algoritmo (Nexa Center, 2018, p. 4). La idea de fondo es que una específica visión de la justicia debe formalizarse con criterios estadísticos, y que estos sean después utilizados para realizar los instrumentos informáticos de los que nos servimos (Santangelo, 2020, p. 1). Para atribuir el calificativo de "justo" al algoritmo, conviene partir de algunas precisiones acerca de la idea de la justicia.

Tal idea de justicia ocupa un lugar central tanto en la ética como en la filosofía jurídica y política. La aplicamos a las acciones individuales, a las leyes y a las políticas públicas. La justicia adquiere diferentes significados en diferentes contextos prácticos, y para comprenderla completamente tenemos que lidiar con esta diversidad. Sin embargo, vale la pena preguntarse si encontramos un concepto central que atraviesa todos estos diversos usos. Tal eje podría ser el contemplado en las Institutas de Justiniano, codificación del derecho romano del siglo VI d. C., donde la justicia se define como "la voluntad constante y perpetua de dar a cada uno lo que le corresponde". Ahora bien, si la justicia tiene que ver con cómo se trata a las personas individuales ("a cada uno lo que le corresponde"), la dificultad radica en identificar qué sea lo que le corresponda a cada uno, ese *suum cuique.* Apunto aquí algunas ideas clave:

- Libertad, oportunidades, recursos, etc. que son potencialmente conflictivos, y apelamos a la justicia para resolver tales conflictos determinando a qué tiene derecho cada persona.
- Aunque la justicia es fundamentalmente una cuestión de cómo se trata a los individuos, también es posible hablar de justicia para grupos, por ejemplo, cuando el Estado asigna recursos entre diferentes categorías de ciudadanos. Aquí, cada grupo se trata como si fuera un individuo separado a los efectos de la asignación.
- Justicia e imparcialidad son un binomio especialmente sensible y que se requiere mutuamente (Barry, 1995). La justicia es lo opuesto

a la arbitrariedad. Requiere que cuando dos casos sean relevantes, deben ser tratados de la misma manera. Ahora bien, cabe plantear la pregunta "¿qué tipo de igualdad requiere la justicia?" (Dworkin, 2003; Sen, 1980); lo que exige la justicia ¿es siempre igualdad, ya sea de trato o de resultado? La justicia requiere la aplicación imparcial en unos casos, pero en otros, como ya advirtió Aristóteles, la justicia también implica la idea de trato proporcional, lo que implica que los destinatarios obtengan cantidades desiguales de cualquier bien en cuestión (Aristóteles, Ética a Nicómaco, Libro V, cap. 3). La igualdad aquí es el principio por defecto que se aplica en ausencia de reivindicaciones especiales que puedan presentarse como razones de justicia.

— El igualitarismo estricto, que exige la asignación de bienes materiales iguales a todos los miembros de la sociedad. La justicia como igualdad y la justicia como merecimiento parecen estar en conflicto, y el desafío es mostrar qué puede justificar la igualdad de trato frente a las desigualdades del merecimiento.
— La justicia distributiva requiere que los recursos disponibles para el distribuidor se comparten de acuerdo con algún criterio relevante, como la igualdad, el merecimiento o la necesidad. Por su parte, el denominado "igualitarismo de la suerte" comprende diversos intentos de diseñar principios distributivos que sean apropiadamente sensibles a las consideraciones de responsabilidad y de suerte (Young, 2011). Los principios distributivos varían en numerosas dimensiones. Varían en lo que se considera relevante para la justicia distributiva (ingresos, riqueza, oportunidades, trabajos, bienestar, utilidad, etc.); en la naturaleza de los destinatarios de la distribución (personas individuales, grupos de personas, clases de referencia, etc.); y sobre qué base debe hacerse la distribución (igualdad, maximización, según características individuales, según transacciones libres, etc.).
— No hay consenso sobre una teoría integral de la justicia (Barry, 1989; Sandel, 2009; Walzer, 2011; Nussbaum, 2006). Algunos autores, se refieren a la misma no como un conjunto de principios sino como una virtud (MacIntyre, 2001). De ahí que, en ocasiones, haya que conformase con teorías parciales: teorías sobre lo que la justicia requiere en dominios particulares de la vida humana. Tales teorías parciales se proyectan sobre la justicia algorítmica y permiten entender las dificultades para lograr un consenso sobre la *fairness* en ML (Cohen, 2011).

La discriminación y afectación de la igualdad constituye uno de los principales obstáculos en la consecución de la justicia, tanto en el Derecho

analógico como en el digital. Toda población está caracterizada por algunos atributos sensibles como la raza, el género, las creencias religiosas, la orientación sexual y otros atributos con respecto a los que se puede desarrollar una discriminación. La *fairness* de un algoritmo es la propiedad de no discriminar respecto a estas características (Galeotti, 2018, p. 75). Al igual que hay dificultades para alcanzar un consenso sobre qué sea la justicia, también las hay para definir un algoritmo equitativo, o una "equidad algorítmica". Cuando el entrenamiento de aprendizaje automático incide y afecta a atributos sensibles relacionados con las personas, la consecución de una equidad e imparcialidad (decisión justa) resulta dificultosa. Y ello porque hay que atender a muchas variables. Basta tomar como referencia la discriminación positiva con respecto a mujeres, personas con discapacidad, personas mayores, que se aplica desde la teoría de los derechos, precisamente porque tratar a todos/as por igual, sin consideración de sus particulares características, desembocaría en una decisión injusta. Esta misma consideración de atributos sensibles debe desarrollarse en la formulación de los algoritmos, en unos casos, para establecer tales discriminaciones positivas y, en otros, para eliminar diferencias.

Si, por ejemplo, lo que se va a controlar es la paridad o el cumplimiento de cuotas para asegurar la representación de los grupos a proteger, la equidad puede medirse contando las personas de los distintos grupos. Sin embargo, cuando se trata de asegurar la equidad en un proceso o decisión, como en un concurso de selección o en un juicio, la medición es mucho más difícil. ¿Cómo medir si el proceso o decisión fue equitativo y no discriminatorio? ¿Se puede aspirar a una completa imparcialidad de la IA? Habría que responder negativamente a esta última pregunta. El objetivo al que se aspira no es la equidad total, sino que, al menos, se establezcan métricas y umbrales de equidad que garanticen la confianza en los sistemas de IA.

Reclamar imparcialidad, equidad y justicia a la IA resulta, para algunos sujetos, más confiable que reivindicarlo de una inteligencia humana, en el sentido de que consideran que tienen una probabilidad mayor de obtener un proceso o una decisión imparcial mediante un sistema automatizado que a través de un juicio humano. Pero ni la ingeniería informática, ni la matemática, ni un entrenamiento estadístico con patrones "imparciales" puede garantizar una equidad total, ya que garantizar la equidad para uno (o un grupo) puede conllevar la injusticia de otro (o de otro/s grupo/s).

2.2. En busca de una definición de fairness algorítmica

De entrada, puede pensarse que el concepto de *fairness* no encaja en el algoritmo ya que este último es diseñado para clasificar, diferenciar,

compartimentar, establecer vencedores y perdedores, especificando qué situaciones conducen a resultados satisfactorios. Los algoritmos de aprendizaje automático supervisados son, de por sí, de naturaleza discriminativa (Stewart, 2020). La razón de tal discriminación estriba en que muchos de los algoritmos se han diseñado precisamente para clasificar los datos en función de las informaciones que aparecen incrustadas en los mismos. A diferencia del "algoritmo de generación" que genera datos según una categoría específica, el "algoritmo discriminatorio" clasifica, divide los datos en diversas categorías, conforme a los criterios que el programador haya establecido, y realiza distribuciones. Los algoritmos son modelados para reconocer características y atributos que en situaciones anteriores han conducido al éxito, no interesándose por lo justo sino replicando modelos y prácticas anteriores. Por ello, para la consecución de un algoritmo justo es imprescindible valorar el significado de elegir un tipo de *fairness* en lugar de otro, en un determinado contexto social (Nexa Center, 2018, p. 5). No es fácil la consecución de este propósito ya que, como explicaré más adelante, no hay una única definición de *fairness* y no se puede implementar simultáneamente más de una definición de *fairness* (Kleiberg *et al.,* 2016). Todo ello permite entender la relevancia que tienen los valores sociales y los conceptos democráticos ligados a las actuales formalizaciones matemáticas de equidad (Barocas *et al.,* 2018; Binns, 2018; Hardt *et al.,* 2016).

La *fairness* es una rama emergente del aprendizaje automático. A diferencia de otros profesionales, como los del ámbito sanitario y los del ámbito jurídico, son todavía escasos los casos en que los ingenieros informáticos reciben capacitación para aprender a considerar el impacto moral de sus acciones de programación y diseño. Sin embargo, nos consta que muchos de estos especialistas ya están trabajando sobre estas cuestiones[5], lo que revela una sensibilidad necesaria tanto con respecto a la consecución de una *fairness* en los algoritmos como de una IA ética. Sus investigaciones se dirigen a mejorar los algoritmos para reducir la dependencia de atributos sensibles y definir indicadores para cuantificar y medir el grado de discriminación.

La *fairness* en la IA trata de garantizar que los modelos de IA no discriminen cuando toman decisiones, particularmente con respecto a atributos protegidos como la raza, el género, el país de origen u otros. Uno de los

5. Una muestra de este interés desde el ámbito de la Ingeniería son los Seminarios impartidos en el ámbito del Programa de Doctorado de Tecnologías Industriales e Ingeniería Civil de la Universidad de Burgos, en 2021, sobre "*Fairness* en la toma de decisiones algorítmica"; otro ejemplo es la asignatura impartida por el Área de Filosofía del Derecho, de la Universidad de Ourense "Dimensión ética y jurídica de la IA",

problemas de medir la *fairness* es que el resultado está específicamente vinculado a la definición que se utilice de la misma. Según las demostraciones matemáticas, estadísticas y de cálculo, es imposible que un modelo sea justo con respecto a varias nociones de equidad simultáneamente. Esto se debe a que tan pronto como se comience a tener en cuenta los problemas para hacer que un modelo sea más justo en función de una definición, surgirán otros problemas que lo harán menos justo de acuerdo con una noción diferente de equidad (Konstantinov, 2022). Si se parte de una equidad como imparcialidad, sin diferenciación por cuotas raciales, de discapacidad o de género, por ejemplo, tal resultado puede ser justo en un determinado contexto, pero no en otro. El desafío es determinar qué umbral de *fairness* es aceptable y que corresponda a un sistema justo.

La *fairness* en los algoritmos y la corrección de resultados injustos y sesgados, constituye un tema crucial en las investigaciones actuales. La equidad algorítmica puede ser informalmente descrita como la probabilidad de ser clasificado en una determinada categoría, la cual, debe ser similar para todos los que exhiban esas características, independientemente de otros rasgos o propiedades. Con el fin de garantizar la equidad algorítmica, cuando se trabaja con conjuntos de datos muy dispares o desequilibrados, los científicos de datos recurren a diversas herramientas. Una de ellas es la calibración (es decir, la comparación de salida real y la salida esperada). Para medir la equidad entre dos grupos (supongamos, en el ámbito de la salud, si se atiende al género, pacientes masculinos y femeninos; o en el ámbito de la justicia penal, si se atiende a la raza, acusados afroamericanos y acusados caucásicos), entonces esta condición de calibración debería mantenerse también simultáneamente para el conjunto de personas dentro de cada uno de estos respectivos grupos (Dignum, 2021). Otra es el umbral de aceptación de las decisiones algorítmicas, que, por su naturaleza, nunca será completamente 'justa'. ¿Cómo medir el mínimo exigible de "justicia algorítmica"? ¿Cómo superar las dificultades que conlleva tanto en su aplicación a nivel individual como a nivel social? Porque la finalidad del Derecho antidiscriminatorio es la de "hacer frente a la exclusión social, a la opresión y a la "subordiscriminación", ofrecer respuestas para abordar graves desigualdades estructurales" (Añón Roig, 2022, p. 43).

Hay varias categorías que se convierten en imprescindibles cuando la filosofía y ML van de la mano, y que permiten entender las dificultades y limitaciones de la construcción de un modelo de predicción perfecto. Elenco sólo las principales:

— Parcialidad: Prejuicio a favor o en contra de una cosa, persona o grupo comparado con otro, generalmente de una manera que se

considera injusta. Acaece cuando un algoritmo produce resultados que contienen prejuicios sistémicos debido a errores y suposiciones en el proceso de aprendizaje automático. Normalmente debido a la recopilación de datos, el muestreo y/o la medición procedimientos.
— Opresión: Un sistema social de barreras que opera institucionalmente e interpersonalmente con la finalidad de desempoderar a las personas en razón de su género, raza, clase, sexualidad, etnia, religión, tamaño corporal, capacidad y/o nacionalidad, u otras características.
— Imparcialidad: Un principio subjetivo de juicio acerca de si una decisión es moralmente correcta o incorrecta. Se trata de un área reciente en el aprendizaje automático que estudia cómo asegurar que los sesgos en los datos, e inexactitudes del modelo, no deriven en modelos que traten individuos desfavorablemente sobre la base de características o atributos sensibles.
— Igualdad: Significa correspondencia entre un grupo de diferentes objetos, personas, procesos o circunstancias que tienen las mismas cualidades en al menos un aspecto, pero no en todos.
— Discriminación: Trato desfavorable de las personas debido a la pertenencia a ciertos grupos demográficos que se distinguen por atributos (supuestamente) protegidos por la ley. Constituye una fuente de injusticia en el aprendizaje automático debido al prejuicio humano (intencional o no) y estereotipos basados en los atributos sensibles.
— Marginación: (Sensible/Protegido): tiene en cuenta aquellas características comúnmente referenciadas y reflejadas en la ley de no discriminación. Por ejemplo, raza, etnia, género, religión, edad, discapacidad, orientación sexual, etc.

Especial atención merecen los algoritmos que se utilizan por parte de las Administraciones públicas para su toma de decisiones (Soriano, 2021). Sus aplicaciones se dirigen a varias áreas tales como, por ejemplo, identificar patrones de evasión de impuestos, filtrar datos de servicios sociales y de salud para priorizar adjudicación de ayudas o subsidios, entre otros. Hay cinco dimensiones que deben de cumplir los algoritmos que se apliquen en este ámbito de la gestión pública: precisión, imparcialidad, explicabilidad, estabilidad y adopción (Dhasarathy *et al.,* 2021): i) la exactitud: normalmente existen muchas medidas posibles y resultados probabilísticos; ii) la explicabilidad: la IA y el ML son más valiosos cuando se utilizan para respaldar-auxiliar, y no sustituyen, la toma de decisiones humanas; deben permitir que los seres humanos comprendan los fundamentos de las recomendaciones del algoritmo —algunos autores ya se refieren al "derecho a la explicabilidad"—; iii) la estabilidad: estimación de la frecuencia con la que

se deben actualizar los modelos, los usuarios deben comprender la velocidad a la que se degrada el rendimiento algorítmico; iv) la adaptabilidad: necesidad de planificar e incorporar enfoques para fomentar la adopción desde el primer día, ya que pueden generar información precisa pero contraria a la intuición debido a la gran cantidad de variables y datos que utilizan, es decir, van contra las heurísticas tradicionales; v) por último, la imparcialidad.

Un primer paso sería el de establecer qué significa la *fairness* en un contexto específico, por ejemplo, cuáles son los grupos vulnerables (si es que los hay), y cuáles serían las métricas para la *fairness*. Dhasarathy et alt. (2021) han señalado algunas formas de medir la *fairness*, tales como las tres siguientes:

- Ceguera voluntaria: Consiste en crear un tipo de ceguera en el algoritmo, de modo que trate a los subgrupos de la misma manera independientemente de las distinciones tradicionales entre ellos, como la raza, el género u otros factores socioeconómicos, como si el algoritmo fuera ciego y, por tanto, clasificara y evaluara los datos de todos los grupos y subgrupos con criterios neutrales. Sin embargo, en lugar de conseguir la *fairness*, este enfoque podría conducir a resultados injustos o causar problemas con los datos de muestra utilizados para entrenar el modelo en sí, acabando por crear un algoritmo que simplemente desconoce sin tener en cuenta la imparcialidad.
- Paridad demográfica o estadística: Para lograr la *fairness* se intenta garantizar la paridad estadística en las decisiones que se habilitan o en los resultados, por ejemplo, seleccionando una proporción igual de personas de los grupos protegidos y no protegidos. Para lograrlo habría que establecer diferentes umbrales para que diferentes grupos aseguren la paridad en los resultados para cada grupo. Por ejemplo, un algoritmo que aplicara diferentes umbrales de puntuación de crédito para diferentes grupos demográficos, a fin de seleccionar la misma proporción de solicitantes de cada uno (mujeres/ hombres). Pero este tipo de algoritmo sólo será efectivo cuando se trabaja con una sola medida de imparcialidad, en este caso, una proporción igual del resultado en la concesión de un préstamo en todos los tipos de género.
- Igualdad predictiva: En este enfoque, no se hace hincapié en el resultado de la decisión, sino en el rendimiento del algoritmo (o exactitud) a través de diferentes grupos. Eso significa, por ejemplo, que las tasas de error o la prevalencia de falsos positivos o falsos negativos para cada grupo son las mismas, al tiempo que se tienen en cuenta las

> variaciones en la población subyacente. Retomado el ejemplo de la concesión de un préstamo bancario, no implica que se apruebe una proporción igual de solicitantes de préstamo entre los géneros, pero el porcentaje de solicitantes aprobados que terminan en incumplimiento (es decir, los falsos positivos) sería el mismo para todos los géneros. Así, no se favorecería o perjudicaría de forma desproporcionada por razón del sexo del solicitante ya que estaríamos cometiendo la misma tasa de errores en nuestra elección (Dhasarathy *et al.,* 2021).

Por su parte, Russel y Norvig (2004) enumeran seis criterios o seis formas de entender qué es la *fairness* o cómo debería ser un algoritmo para ser considerado imparcial —criterios que, cada uno por separado, resultan razonables, pero difícilmente pueden satisfacerse al mismo tiempo (Barocas, Hardt y Narayanan, 2018)— que resultan ilustrativos para comprender las cuestiones a las que aludiré más adelante:

- Imparcialidad individual ('*individual fairness*'): requisito de que cada individuo sea tratado de forma similar a otros individuos con independencia de sus atributos, características o clase (género, raza, edad, etc.)
- Imparcialidad de grupo ('*group fairness*'): requisito de que dos clases sean tratadas de forma similar lo cual sea demostrable estadísticamente (por ejemplo, que individuos de raza blanca y de raza negra sean tratados de forma similar).
- Imparcialidad a través de la inconsciencia ('*fairness through unawareness*'): requisito de que si eliminamos atributos referentes a aspectos sensibles, como el género o la raza, de un conjunto de datos ('*dataset*'), el algoritmo no pueda discriminar con base en esos atributos. Aunque el algoritmo desconozca esos atributos, hay otras muchas variables, variables 'proxy', que dan 'pistas' del valor de esos atributos que se pensaba habían quedado eliminados del conjunto de datos. El algoritmo de contratación de Amazon (Kraus, 2018), aunque no se indicara el género, por otras informaciones que las solicitantes de empleo indicaban en su *curriculum vitae*, detectaba que se trataba de mujeres y las discriminaba, penalizando su contratación, razón por la que Amazon retiró tal programa de contratación de personal.
- Igual resultado ('*equal outcome*'): requisito de que cada clase demográfica obtenga los mismos resultados, que es lo que se conoce también como paridad demográfica ('*demographic parity*'). Por ejemplo, en el caso de solicitud de préstamos bancarios, la paridad demográfica garantizaría que se concede el mismo volumen de préstamos a

hombres y mujeres. Nótese que la paridad demográfica no asegura que se cumpla otro de los criterios, como es el de la imparcialidad individual.

- Igual oportunidad ('*equal opportunity*'): requisito, también conocido como equilibro ('*balance*'), que significa, siguiendo con el ejemplo del préstamo, que si un individuo tiene capacidad de devolver el préstamo debe serle concedido independientemente de que se trate de hombre o mujer. Nótese de nuevo que, cumplir este criterio, puede llevar a no cumplir el de igual resultado si hubiera diferencias en la capacidad para devolver el préstamo entre hombres y mujeres.
- Igual impacto ('*equal impact*'): parecido al anterior pero yendo un poco más allá, pues pide que la utilidad ('*utility*') esperada ante una situación como la del préstamo sea igual independientemente de la clase a la que se pertenezca, es decir, se valoran tanto los beneficios de una predicción verdadera, como los perjuicios de una predicción falsa.

Si bien los seis criterios anteriormente enunciados son suficientemente ilustrativos para comprender las distintas métricas y combinaciones de la *fairness*, hay otras propuestas. Una de ellas es la expuesta en un artículo que dos investigadores canadienses, Julia Rubin y Sahil Verma (2018), presentaron en un workshop internacional sobre *fairness* del software, en el que analizan hasta veinte definiciones de *fairness*. El estudio de Rubin y Verman aplica los veinte conceptos de *fairness* de ML a un supuesto concreto como es la predicción de la calidad crediticia de un solicitante. Para la muestra, analizaron 1000 registros de solicitantes de crédito alemanes, a partir de veinte datos o atributos de cada solicitante (saldo de cuenta; meses del crédito; historial crediticio; destino del crédito; monto del crédito; saldo cuenta de ahorro/inversión; tiempo en el actual empleo; porcentaje de mensualidad/ingreso; sexo y estado civil; codeudores y/o aval; tiempo en residencia actual; activos o propiedades; edad; otros pagos a plazos; casa propia o en renta; otros créditos del mismo banco; tiempo de empleo; dependientes económicos; teléfono a su nombre, si es trabajador extranjero), así como la calidad crediticia —buena o mala en función de si pagó o no su crédito—. A partir de esos datos, los investigadores desarrollaron un modelo de clasificación que, a partir de esos veinte atributos, predice la calidad crediticia de un solicitante y, por tanto, permite a la entidad bancaria decidir a quién concede o a quién deniega un crédito.

Más allá de la "justicia predictiva" que este tratamiento pueda implicar, se trataba aquí de evaluar si las predicciones del algoritmo eran equitativas para hombres y mujeres no solteros. Aplicando el algoritmo a este conjunto

de datos y a varias métricas (paridad estadística, paridad predictiva, balance de falsos positivos y balance de falsos negativos, equidad de exactitud condicionada, equidad de exactitud general y equidad de tratamiento) el resultado fue que los hombres no solteros tienen una probabilidad mayor de obtener una predicción positiva (0.81) que las mujeres no solteras (0.75), por lo que no hay paridad estadística. A juicio de los investigadores, este resultado evidencia un sesgo, por el cual es más fácil que hombres con mala calidad crediticia reciban una buena predicción. Por otro lado, cuanto más baja es la probabilidad pronosticada de una buena calidad crediticia, aumenta más la posibilidad de un error. Con esta información pueden analizarse los casos donde se presentan diferencias y decidir entre varios cursos de acción: modificar los atributos utilizados para la predicción, darles un peso diferente, y/o intentar el uso de un algoritmo distinto. Como advierten Carey y Wu (2022), la mayoría de las métricas propuestas de *fairness* de aprendizaje automático se basan en medidas estadísticas.

Sin embargo, los fundamentos estadísticos no brindan equidad a nivel individual, o incluso de subgrupo, sino que sólo aportan garantías significativas al miembro "promedio" de un grupo marginado. Además, muchas medidas estadísticas se oponen entre sí. Ello aconseja buscar unos fundamentos que la sustenten.

3. Fundamentos iusfilosóficos para una teoría de la *fairness* del algoritmo

La "justicia" proporciona un conjunto de estándares conforme a los cuales se adjudica "justamente". Ahora bien, ese "justamente" puede significar de forma equitativa, imparcial, igualitaria, de forma diferenciada, u otras. Basta recordar que los criterios de justicia, a lo largo de la historia, han sido diversos ("justicia es lo que Dios quiere", "justicia es lo que corresponde a la naturaleza humana", "justicia es dar a cada uno lo que le corresponde", "justicia es dar a cada uno lo que merece", "justicia es tratar a todos por igual", entre otras). Sin recurrir a la conclusión a la que llegaba Kelsen sobre qué es justicia ("No obstante, ahora como entonces carece de respuesta"), debe advertirse que su concepción está influida y modulada en función de la corriente a la que se adscriba. Como ha subrayado Bellver, al digitalismo lo alienta una demanda de justicia y emancipación, que busca acabar con formas de discriminación muy arraigadas y procurar la igualdad efectiva entre todos los seres humanos. Sin embargo, se sustenta sobre una base filosófica que niega la inteligibilidad de la realidad y la condición teleológica de la existencia humana proponiendo, en su lugar, como única

guía para orientar la vida humana, la hegemonía del deseo individual (Bellver Capella, 2021, p. 21).

Es de la mayor importancia identificar qué enfoque o concepción de la justicia se elija para presidir la *fairness* que acompañará al algoritmo, porque desplegará sus efectos sobre la calidad de nuestra vida democrática, y porque en la actualidad dependemos de los procesos de toma de decisiones mediante el uso de la IA, a través de los cuales se realizan adjudicaciones (Santangelo, 2018). El algoritmo decidirá si los honores deberán ser conforme al merecimiento de los individuos; la educación superior, conforme al talento; la riqueza conforme a la habilidad y la suerte en el merecimiento, y así de forma sucesiva. Pero ¿estos criterios de distribución son satisfactorios? y, sobre todo, ¿son justos? La elección entre unos objetivos correctivos y/o distributivos ya plantea un dilema. El contraste entre justicia correctiva y distributiva puede remontarse a Aristóteles. El objetivo rector de la justicia correctiva se refiere a una relación bilateral entre el malhechor y su víctima, enfatizando el remedio que restituye a la víctima al estado que tenía antes de que ocurriera el comportamiento ilícito. En cambio, el objetivo distributivo de la justicia implica una relación multilateral, y formula la justicia como principio para repartir bienes de diversa índole a los particulares. Si bien la justicia correctiva aparece con más frecuencia en las prácticas legales, la equidad en el aprendizaje automático se limita en gran medida a la consideración que tienen objetivos distributivos de justicia, como pueda ser la distribución de oportunidades de admisión.

Basta tomar como ejemplo, el reciente Anteproyecto de Ley Orgánica del Sistema Universitario (LOSU), aprobado por el Gobierno en junio de 2022, conocida como "Ley Subirats". El texto, que derogará la LOU de 2001 reformada en 2007, contiene algunos preceptos que llevan a preguntarse sobre su equidad e imparcialidad, hasta el punto de que algunos Diarios han abierto sus titulares como textos de este tenor "El Gobierno aprueba la Ley Subirats que prioriza contratar a mujeres y becar a extranjeros" (Diario *El Mundo,* 21.06.2022). Vamos a plantear la hipótesis de que, para llevar a cabo tales contrataciones en la Universidad, se utilizara un programa de IA que recibiera, clasificara, valorara las solicitudes y resolviera automáticamente a quién contratar. El algoritmo del programa debería de incorporar una discriminación positiva para las mujeres. No se trata aquí de evitar sesgos por razón de género, sino por el contrario, se deberían de introducir para que, como establece la ley, cuando haya dos candidatos en "igualdad de condiciones de idoneidad", tengan "preferencia" para ser contratadas "las personas del sexo menos representado" en el cuerpo docente o categoría de que se trate. Es decir, las mujeres tendrán prioridad frente a los hombres a la hora de trabajar en las Universidades públicas. En el Derecho analógico,

tal requisito ya ha generado controversia porque si dos personas compiten, no se puede tomar el sexo como criterio. En el ámbito digital, tendría que reflejarse esta misma discriminación positiva. De ahí que el programador debería de incorporar las exigencias legales de discriminación positiva a la hora de diseñar el software de contratación.

3.1. Algunos marcos teóricos sobre la justicia para entender la justicia algorítmica

Voy a detenerme, entre otros, en cuatro marcos teóricos bajo los cuales se puede entender la justicia: el utilitarismo, el contractualismo, el comunitarismo y el igualitarismo (Zeyu *et al.,* 2022):

— *La perspectiva utilitaria de la justicia.* El utilitarismo tiene como objetivo maximizar el bienestar, y producir la mayor cantidad (en términos de las utilidades agregadas) de bien para el mayor número de individuos. El utilitarismo juzga los resultados sumando los niveles de utilidad y no tiene una preocupación independiente sobre cómo se distribuye esa utilidad entre las personas. La evaluación es solo con respecto a las consecuencias sin ninguna consideración sobre cómo se llega a tales resultados. Es decir, la justicia o la equidad debe tomar la forma de beneficios/cargas, los medios para obtener la felicidad en lugar de la felicidad/infelicidad misma (Zeyu *et al.,* 2022).

— *La perspectiva contractualista de la justicia.* Los filósofos contractualistas abordan la justicia mirando para principios (hipotéticos) en forma de acuerdos a los que se comprometen las instituciones y los individuos. John Rawls presenta el escenario donde los individuos saben que sus "concepciones del bien" son en general diferentes, pero al mismo tiempo, la concepción del bien de cada individuo se coloca detrás de un ""velo de ignorancia". Su concepción de justicia entendida *como equidad,* ya había sido propuesto en *A Theory of Justice* (1971), y se reformula en *Justice as Fairness* (2001), dando forma a una teoría de la justicia a partir de la idea de un contrato social. Los dos principios de justicia (igual libertad para todos y principio de la diferencia) sustentan una determinada concepción de la justicia. Las desigualdades económicas y sociales tienen que satisfacer dos condiciones: en primer lugar, tienen que darse en condiciones de igualdad equitativa de oportunidades; en segundo lugar, las desigualdades deben redundar en beneficio de los miembros menos aventajados de la sociedad (el principio de diferencia). Dicha concepción se explica desde valores polí-

ticos y no desde una doctrina moral, religiosa o filosófica. Por su parte, T. M. Scanlon explica la idea de justicia en términos de "lo que nos debemos unos a otros" y la presenta como un acuerdo general que ningún individuo, informado y no forzado, podría rechazar de forma razonable (Scanlon, 1998; Scanlon, 2000).

A juicio de Santangelo, la mayor parte de quienes trabajan sobre *fairmess* algorítmica adoptan una visión típicamente liberal del concepto de justicia, centrada sobre la tutela de los derechos individuales y sobre la probabilidad que cada uno tiene de que sus derechos sean respetados por los instrumentos informáticos y menos, por una concepción de la justicia de tipo distributivo, es decir, más en la línea de la propuesta de Rawls, más típica de la social-democracia. Del lado de los liberales, además de Rawls, se puede citar a autores como Thomas Nagel y Ronald Dworkin.

— *La propuesta de justicia de comunitaristas,* como Alasdair MacIntyre, Michel Sandel, Charles Taylor y Michel Walzer. Así, como explica Sandel (2018, 296), la justicia "no solo trata de la manera debida de distribuir las cosas, sino que trata también de la manera debida de valorarlas". Los liberales, herederos de Locke, Kant y Mill, comparten la misma defensa de la libertad de conciencia, respeto por los derechos del individuo y la misma desconfianza con respecto a la amenaza de un Estado paternalista. Por su parte, los comunitaristas, tomando sus raíces del aristotelismo, de la tradición republicana renacentista y de la hermenéutica contemporánea, comparten una desconfianza hacia la moral abstracta, una simpatía hacia la ética de la virtud y una concepción de la política inclinada hacia la historia y las tradiciones.

Este planteamiento entre liberales y comunitaristas se puede circunscribir a unas antinomias esenciales, como individuo-comunidad, derechos-virtudes, justo-bien. Las respectivas teorías de la justicia de Walzer (teoría de la complejidad) (2001) y de Rawls (justicia como equidad) son un exponente de esa variedad de concepciones sobre la justicia y permiten comprender la dificultad que tiene este tema, ya como punto de partida, a la hora de plasmarlo en un algoritmo. A juicio de algunos autores, no es tanto el debate entre liberalistas y comunitaristas pues, en ambos entornos hay liberales, sino que se trata de un debate entre la concepción moderna del hombre, individualista, racionalista y universalista *versus* antropologías sociales, históricas, hermenéuticas y contextualistas (Borges de Macedo, 2001, p. 28). Walzer desarrolla su planteamiento sobre la justicia distributiva, de tipo particularista, confrontando la tesis de Rawls, según la cual existiría un criterio distributivo universal para todo tipo de bienes. Los tres principios

de distribución que señala (el intercambio libre, las necesidades y el mérito) se inscribe en una defensa de la libertad sin renunciar a la igualdad.

Por regla general, los ciudadanos compiten en cuanto a la demanda de libertades, oportunidades, recursos y modos de tratamiento. El concepto y criterio de justicia está influido por elementos socio-institucionales, a los que conviene hacer referencia. Rawls, al ocuparse sobre la "justicia" aplicada a las principales instituciones sociales, señala que, incluso las personas que no están de acuerdo con lo que exige la justicia pueden estar de acuerdo en que necesitan algunos estándares para cumplir esta función —determinar cuáles son los derechos y deberes de las personas y cómo debe ser la distribución de los beneficios y de las cargas por el hecho de vivir juntos— (Rawls, 2006). Ahora bien, no todos los teóricos políticos definen la justicia en términos rawlsianos. Así, el velo de la ignorancia que, a juicio de Rawls, evita que el pueblo en la posición original conozca su identidad y sus características personales y sociales, no encajaría con el conocimiento que hay que tener sobre las injusticias del mundo real y las divisiones sociales (por ejemplo, de raza y de género) y que informan y ayudan a conformar nuestro razonamiento sobre la justicia y, por ende, ser tomadas en consideración para lograr la *fairness* del aprendizaje automático.

— *La perspectiva igualitaria de la justicia.* El igualitarismo pretende establecer algún tipo de igualdad. En cierta medida, la igualdad podría actuar por defecto cuando comprendemos intuitivamente la idea de equidad y justicia. Una pregunta natural que enfrenta el igualitarismo es cómo hacer que la idea de equidad sea como la igualdad más específica y razonable en diferentes contextos. La distribución igualitaria (de oportunidades) como un punto de partida, permite desviaciones de la línea de base de igualdad si tales desviaciones resultan de elecciones responsables de individuos; el igualitarismo de la suerte, como una llamada a la responsabilidad, agrega una restricción adicional de que en las desigualdades resultantes la suerte bruta debe ser restringida (Miller, 2001).

Las estructuras sociales son poderosos agentes de desigualdad, discriminación, prejuicios y parcialidad. Como ha apuntado Young, las personas ocupan diferentes posiciones dentro de una "sociedad estructurada", cada una de ellas con sus correspondientes expectativas, ventajas, y desventajas. Las estructuras sociales se convierten en lugares de injusticia cuando empoderan sistemáticamente a las personas en algunos posiciones al desempoderar a otros (Young, 2000) —como en el caso de los hombres en el mercado laboral cuando, en el caso ya citado, al aplicarse el algoritmo que utilizaba Amazon para la contratación laboral, eran empoderados porque las mujeres estaban desempoderadas—. Un algoritmo puede aplicar formas

de acción positiva o acciones afirmativas para compensar de manera directa algunas categorías desaventajadas, especialmente, algún grupo desventajado, actuando de este modo como una forma de "acción positiva" (Calderes, 2009; 2010).

En esta línea, Crenshaw (1991) argumenta que las personas pueden experimentar la opresión de manera diferente en función de su superposición identidades, lo que denomina como "interseccionalidades". Por ejemplo, una mujer de raza negra y pobre puede experimentar opresión no experimentada por otras mujeres, o por otras personas de raza negra o por otras personas que dispongan de una economía saneada. Instituciones, estructuras de poder, sistemas políticos, contribuyen a incrementar o reducir las injusticias. De ahí que todos estos factores deban ser tomados en consideración para configurar la *fairness* algorítmica. Ahora bien, esto lleva a plantearse si lo que en realidad se está pidiendo al algoritmo es que sea capaz de procesar de forma adecuada el sumatorio de atributos sensibles de un individuo para atribuirle la discriminación positiva que le correspondería por cada característica.

3.2. El sustrato socio-técnico y jurídico-político para identificar una métrica de fairness

La dificultad en el mundo analógico de alcanzar la imparcialidad no se resuelve en el mundo digital simplemente con los procedimientos matemáticos y algorítmicos de varios populares métodos de *fairnes*s ML estadísticos y causales. La igualdad estadística, e indicadores como la igualdad de oportunidades y la probabilidad de compensación, por ejemplo, son más que fórmulas matemáticas y están sustentadas en unos fundamentos filosóficos, sociológicos, jurídicos y políticos.

Carey y Wu (2022) son algunos de los investigadores que se han preocupado por explicar los fundamentos filosóficos subyacentes y pensamientos jurídicos que sustentan los algoritmos. La identificación de una métrica de *fairness* debe tomar en consideración los impactos sociales que pueden surgir una vez se implementara ya que, como he indicado antes, los algoritmos y su proceso decisional impacta en numerosos aspectos de la vida de las personas. Por ello, los algoritmos no pueden limitarse a constructos matemáticos, estadísticos y de programación, sino que deben de tener presente la sociología, la política, la filosofía, el Derecho, y la justicia que subyace a los mismos. Ambos autores señalan varias perspectivas filosóficas y fundamentos filosóficos de la IA para justificar las varias métricas de igualdad (la igualdad de oportunidades, y los efectos legales, como el impacto dispar y

el trato desigual). Haré una referencia breve a las mismas ya que todas ellas ayudan a calibrar las métricas de la *fairness*.

Tomando como base la justicia distributiva, el ideal filosófico de la igualdad de oportunidades (*Equality of Opportunity* –EOP–) admite tres acepciones: EOP formal, EOP sustantiva y EOP suerte-igualitaria. La igualdad de oportunidades es un ideal político que se opone a la jerarquía asignada al nacer (casta), pero no a la jerarquía misma. Es un principio que dicta cómo deben ser las posiciones u oportunidades deseables distribuidas entre los miembros de una sociedad. Presenta preguntas morales que los profesionales del aprendizaje automático deben responder para guiar la construcción de un sistema que tiene ideales de *fairness* que satisfagan sus valores deseados. Las tres distintas concepciones de la EOP interpretan la idea de competir en igualdad de condiciones de diferentes maneras.

i) EOP formal enfatiza que cualquier posición deseable en una sociedad, o más concretamente, por ejemplo, una oferta de trabajo esté disponible y abierta a todos. La distribución de estos puestos deseables sigue de acuerdo con las calificaciones relevantes del individuo, y en este escenario, siempre gana el más cualificado. Podría implementarse en ML a través de la "ceguera", es decir, las métricas formales basadas en EOP eliminan cualquier atributo de marginación irrelevante, como la raza o el género. Sin embargo, aunque la EOP formal tiene la ventaja de otorgar puestos en función de las calificaciones reales de una persona, excluyendo la información de marginación irrelevante, no intenta corregir los privilegios arbitrarios. Por ejemplo, en la tarea de predecir el rendimiento académico de los futuros estudiantes para su uso en las decisiones de admisión a la Universidad, las personas que pertenecen a grupos marginados, como los estudiantes de otras razas o del colectivo LGBTB+, son desproporcionadamente afectados por los desafíos de la pobreza, el racismo y la discriminación.

ii) EOP sustantiva, va un paso más allá que la formal, y se dirige a que todos los individuos tengan la misma oportunidad de obtener cualificaciones. Su objetivo es dar a todos una oportunidad justa de éxito en una competencia. Por ejemplo, hacer que todas las actividades extracurriculares y las oportunidades que brindan los estudios universitarios estén también disponibles para todos los estudiantes independientemente de su riqueza o estatus social. Esta métrica estaría en la línea de la propuesta rawlsiana que establece que todos los individuos, sin importar cuán ricos o pobres nazcan, deben tener las mismas oportunidades para desarrollar sus talentos, de manera que personas con los mismos talentos y motivación tengan las mismas oportunidades. En una fairness ML, el EOP sustantivo a menudo

se implementa a través de métricas como la paridad estadística y las probabilidades igualadas, de forma que el talento y la motivación se distribuyen por igual entre las subpoblaciones.

iii) EOP igualitaria de la suerte (Suerte-EOP igualitario**)** impone que el resultado de una persona debe verse afectado sólo por sus elecciones, no por sus circunstancias. Por ejemplo, en el caso de que un estudiante rico se esfuerce en sus estudios, no implica que debería ser penalizado por gozar de buena posición económica. Lo que esta métrica defiende es que la base de la decisión no sea el privilegio o circunstancia arbitraria sino la habilidad relevante.

Sin embargo, como puede apreciarse, esta EOP no se ajustaría a lo establecido en la exigencia legal del proyecto de LOSU de contratación en la Universidad priorizando a las mujeres, ya que la fundamentación jurídica que subyace en la propuesta legal es la discriminación positiva por razón de género y de nacionalidad (por ejemplo, al incentivar la concesión de becas a los extranjeros).

Por último, con relación a la diferencia entre impacto dispar (desigual) (*disparate impact*) y tratamiento dispar (desigual) (*disparament treatment)*, Xiang y Raji (2019) explican que el impacto dispar ocurre cuando los miembros de una clase marginada se ven afectados de manera negativa más que otros al usar una política o regla formalmente neutral. En otras palabras, es una discriminación no intencional o indirecta. El impacto dispar, en sí mismo, no es ilegal. Por ejemplo, la discriminación indirecta en el empleo no es ilegal si puede estar justificada por un "objetivo legítimo" o un requisito profesional genuino y/o necesidad empresarial, como en nuestro ordenamiento pueda ser la cuota para las personas con discapacidad. A diferencia del impacto dispar, el trato desigual ocurre cuando un individuo es tratado intencionalmente diferente basado en su pertenencia a una clase marginada. Sería el caso de que, a miembros de una raza, género o grupo étnico, se les niegue el mismo empleo, promoción u otras oportunidades de empleo que hayan estado disponibles para otros empleados o solicitantes. Habría que dilucidar si las acciones de esa discriminación han estado motivadas por una intención discriminatoria.

De lo expuesto se deduce, en primer lugar, las dificultades que plantea la discriminación algorítmica —lo que ha llevado a formular propuestas como la de reconducir la discriminación hacia modalidades distintas a la discriminación indirecta, tales como la discriminación por asociación y la discriminación interseccional—; en segundo lugar, que la identificación entre discriminación directa e indirecta de la legislación y jurisprudencia

europea con el *disparate treatment* y el *disparate impact* de la jurisprudencia norteamericana, no es pacífica (Añón Roig, 2022, 36).

4. La imprescindible correlación entre la *fairness* artificial y la *fairness* legal

De lo expuesto, puede apreciarse que las categorías de imparcialidad, equidad, igualdad de oportunidades, discriminación positiva, tratamiento desigual, impacto desigual, y tantas otras, se vienen utilizando en la IA cuando se aplica al Derecho. El ámbito digital no puede quedar desligado del ámbito legal, ya que la IA se diseña para aplicarla a unos contextos, los cuales, siempre van a estar regidos (y limitados) por el Derecho (y, también, por la ética). Tal correlación tiene una trascendencia jurídica, que se proyecta en la *fairness*. Para demostrar que se alcanza un determinado umbral de *fairness* algorítmica, las definiciones de la misma (y las métricas y criterios que se elijan) deben corresponder con precisión a sus contrapartes legales para que, por ejemplo, en una decisión de la Administración pública sobre asignación de viviendas sociales, o, en el caso español, en la ya conocida controversia acerca de cómo el algoritmo asigna ayudas públicas relativas al bono social eléctrico, se establezca quién es el responsable (un algoritmo no tiene responsabilidad). La tensión entre *fairness* ML y *fairness* legal no tiene una relevancia meramente conceptual, sino que se proyecta en el desarrollo ordinario de la vida social y, en consecuencia, también en las acciones que llegan a los tribunales. Considerarse objeto de discriminación en la obtención de una beca de estudios para acceder a la Universidad, en la asignación de una vivienda social, en una contratación laboral, en la percepción de ayudas para el pago del gasto energético, en la concesión de libertad provisional en el sistema penal, en la contratación como profesora en una Universidad, exige también, clarificar los fundamentos filosófico-jurídicos que subyacen al diseño de esos algoritmos que dan lugar a tales decisiones: igualdad, imparcialidad, equidad, igualdad de oportunidades, no-discriminación, opresión, en el ámbito legal, y combinarlas con las métricas con que se trabaja en el ámbito del ML: igualdad predictiva, paridad demográfica, trato desigual, impacto dispar, etc.

Causalidad e intencionalidad son, por ejemplo, dos requisitos que ayudan a establecer la *fairness* legal y que son muy difíciles de encontrar en la *fairness* de ML. En primer lugar, para sustentar que ha habido una discriminación hay que establecer, de manera necesaria, el nexo causal. Es decir, la discriminación (sesgo) deberá demostrarse a través del motivo, intención demostrada de exclusión y causalidad, no simplemente por los resultados.

Asimismo, quienes desarrollan los algoritmos deben adoptar las debidas cautelas y medidas particulares de sesgo, para que las métricas se ajusten a lo que los operadores jurídicos acepten como evidencia de discriminación o de no-discriminación. En segundo lugar, no se trata de reemplazar al decisor humano por un decisor automático (algoritmo). En un trato desigual o dispar, no se trata sólo de verificar el resultado sino la intencionalidad (en el caso de una contratación, que para ese puesto concreto se necesite, por ejemplo, que sean de un determinado sexo; en la LOSU, no dar prioridad a las mujeres en la contratación, en "igualdad de condiciones de idoneidad", daría lugar a una vulneración de la norma). La intención es una característica humana y un algoritmo difícilmente podrá valorar la intencionalidad (Xiang, Raji, 2019).

La variedad de conceptos de *fairness* artificial y las dificultades para hacerlo coincidir con los fundamentos que sustentan la *fairness* legal, hacen que algunos autores recurran a la frase de "aversión a los algoritmos" en el sentido de que describen su actitud de oposición a que se usen los algoritmos como instancia decisora, principalmente a la hora de adoptar decisiones sobre determinadas políticas públicas (Dietvorst *et al.,* 2015).

5. Reflexiones finales

Se debate sobre si una sociedad digital puede (y debe) ser ética. El algoritmo no tiene por qué tener una ética, ya que son las personas (empresas, programadores, Administraciones, usuarios) quienes deben de hacer viable que el algoritmo responda a una ética. Además de la ética, se necesita el Derecho y la justicia, porque no basta con las buenas intenciones. No hay que dejarse envolver por los grandes discursos y relatos que presentan al algoritmo como el nuevo *logos.* La tecnoética no consiste en prescripciones éticas individuales sino globales y debe de ir de la mano del tecnoderecho. La resistencia de los investigadores a someterse a normas jurídicas porque se considera que coartan y limitan los avances de la ciencia, se falsean, porque la ciencia también busca unos valores. La respuesta ética es precautoria, preventiva; la respuesta jurídica es reactiva, se articula mediante reparaciones, y postula la adopción de medidas para valorar el daño, probarlo y, en su caso, repararlo. No hay dos vertientes de una misma realidad jurídica, sino que la realidad digital se basa en principios diferentes.

Se cuestiona si el algoritmo debe ser justo. Se repite el relato de que el ámbito del algoritmo no es para los juristas, que los algoritmos (sus diseñadores y programadores) no están interesados por los principios, ni por pacificar, ni por proteger derechos fundamentales. Parece que, en el ámbito

digital, la categoría fundamental es la protección de datos y que, jurídicamente, todos los constructos deben de girar sobre tal propósito. Como ya hemos señalado en la introducción, nada más lejos de la realidad, ya que sí hay una preocupación por parte de los especialistas en lograr un algoritmo que se ajuste a la *fairness.* Basta consultar la abundante bibliografía que han generado investigadores no sólo de la filosofía jurídica y política sino, sobre todo, de la ingeniería informática y de las matemáticas, que están investigando sobre la *fairness*. Si se considera que del positivismo jurídico se ha transitado hacia un positivismo algorítmico, y que las decisiones se adoptan cada vez más a través de los algoritmos, la exigencia de justicia que se lleva a cabo con respecto a la legislación, debe hacerse también para un instrumento que, en un lenguaje informático, está aplicando también esas leyes.

La falta de consenso sobre qué sea la *fairness* y las respectivas fundamentaciones iusfilosóficas que subyacen como consecuencia de diversas concepciones de la justicia (utilitarista, contractualista, comunitarista, igualitarista), las vertientes que la no-discriminación exigen para proyectarse en las políticas públicas, la interacción sociedad-democracia con la concepción de *fairness*, ponen sobre la mesa la imprescindible correlación que debe de existir entre la investigación que lleven a cabo los expertos en ciencias de la computación y los juristas. El diálogo interdisciplinar y la colaboración entre ambos campos de conocimiento son imprescindibles para avanzar en los usos de la IA en el Derecho en concreto, y en la vida social en general.

La fairness algorítmica debe plantearse como un problema sociotécnico, en lugar de simplemente técnico. Individuos y grupos subrepresentados y discriminados en la vida real, deben ser tomados en consideración en el mundo digital. El diseño de un algoritmo debe ir más allá de la simple consecución de unos logros técnicos, de unos resultados en los centros de computación y en la investigación, porque va a desplegar sus efectos en las personas (bien sea como equidad grupal o como equidad individual), por lo que resulta necesario que se centren en eliminar la discriminación también a través de unos algoritmos imparciales, equitativos y justos. Precisamente, por las insuficiencias de los fundamentos estadísticos, se genera la acuciante necesidad de recurrir a fundamentos iusfilosóficos. A pesar de la introducción de vectores correctores y técnicas de calibración y ajuste, la *fairness* algorítmica es de difícil consecución —lo cual incentiva a seguir trabajando y mejorando los logros hasta ahora obtenidos—. La aspiración no debe limitarse a la consecución de un algoritmo "explicable" sino también "justo", que incorpore la *fairness*, y que sea capaz de conseguir aunar el mundo analógico y el digital en el ámbito jurídico.

Referencias bibliográficas

Anderson, Elizabeth S. (1999). What Is the Point of Equality? *Ethics. International Journal of Social, Political, and Legal Philosophy*, 109(2), 287-337.

Añón Roig, María José (2022). Desigualdades algorítmicas: conductas de alto riesgo para los derechos humanos. *Derechos y Libertades*, Época II, 47, 17-49.

Barocas, Solow & Selbst, Andrew D. (2016). Big data's disparate impact. *California Law Review* 104 (3), 671-732.

Barocas, Solon; Hardt, Moritz & Narayanan, Arvind (2018). *Fairness and Machine Learning*. fairmlbook.org.

Barrio Andrés, M. (ed.) (2019). Legal Tech. La transformación digital de la abogacía. Wolters Kluwer.

Barry, Brian (1989). *Theories of Justice*. Hemel Hempstead: Harvester-Wheatsheaf.

Barry, Brian (1995). *Justice as Impartiality*. Oxford: Oxford University Press.

Bellver Capella, Vicente (2021) "Transhumanismo, discurso transgénero y digitalismo: ¿exigencias de justicia o efectos del espíritu de abstracción?, *Persona y Derecho*, vol.84, pp. 17-53. https://revistas.unav.edu/index.php/persona-y-derecho/article/view/40651

Binns, Reuben (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, 81, 149-159.

Borges de Mazedo, U. (2001). A ética do futuro. En: *A presença da moral na cultura brasieira. Ensaio de Ética e Historia das Idéias no Brasil*, Londrina, UEL.

Buchanan, Bruce G. & Headrick, Thomas E. (1970). Some Speculation about Artificial Intelligence and Legal Reasoning, *Stanford Law Review*, 23(1), 40-62.

Campione, Roger (2021) *La plausibilidad del Derecho en la era de la Inteligencia Artificial. Filosofía carbónica y filosofía silícica del derecho*. Madrid: Dykinson.

Carey, Alycia N. & Wu, Xintao (2022). The fairness field guide: perspectives from social and formal sciences, arXiv:2201.05216v2 [*cs.AI*] 8.

Cárdenas Krenz, Ronald (2021). ¿Jueces robots? Inteligencia artificial y derecho ¿Judges robots? Artificial intelligence and law. *Justicia &Derecho*, 4, 1-10.

Casadei, Thomas & Pietropaoli, Stefano (2021). Intelligenza artificiale: fine o confine del diritto? En: Cassadey, Thomas; Pietropaoli, Stefano (a cura di), *Diritto e Tecnologie Informatiche* (pp. 219-232). Milán: Wolters Kluwer.

Caton, Simon & Haas, Christian (2020). Fairness in Machine Learning: a survey, *arXiv preprint. arXiv:2010.04053*, 1-33.

Cohen, G.A. (2011). Fairness and Legitimacy in Justice, and: Does Option Luck ever Preserve Justice? En: *On the Currency of Egalitarian Justice and Other Essays in Political Philosophy*, edited by Michael Otsuka. Princeton NJ: Princeton University Press.

Courtland, Rachel (2018). Bias detectives: the researchers striving to make algorithms fair. *Nature,* 558, 357-360.

Crawford, Kate (2021). *Atlas of AI. Power, Politics and the Planetary Costs of Artificial Intelligence*.

Crenshaw, Kimberle (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color, *Stanford Law Review*, 43(6) 1241-1299. https://doi.org/10.2307/1229039

Chouldechova, Alexandra (2016). Fair Prediction with Disparate Impact: A study of Bias in recidivism prediction instruments, Big Data.

Dastin, Jeffrey (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/amazoncom-jobs-automation-idUSL2N1VB1FQ

Dhasarathy, A., Jain, S. & Khan, N. (2020). *When governments turn to AI: Algorithms, trade-offs, and trust.* McKinsey & Company. https://www.mckinsey.com/industries/public-and-social-sector/our-insights/when-governments-turn-to-ai-algorithms-trade-offs-and-trust.

De Asís Roig, Rafael (2014). *Una mirada a la robótica desde los derechos humanos*. Madrid: Dykinson.

De Asís Roig, Rafael (2022). *Derechos y tecnologías*. Madrid: Dykinson.

Dietvorst, Berkeley J; Simmons, Joseph P. & Massey (2015). Cade. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144 (1) 114-126. doi: 10.1037/xge0000033.

Dignum, Virginia (2021). The Myth of Complete AI-Fairness, *cs.CY*, 1-6.

Dwork, Cynthia; Hardt, Moritz; M., Pitassi, Toniann; Reingold, Omer & Zemel, Rich. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-226.

Dworkin, Ronald (2003). *Virtud soberana: la teoría y la práctica de la igualdad*. Barcelona: Paidós.

Eubanks, Virginia (2021). La automatización de la desigualdad. *Herramientas de tecnología avanzada para supervisar y castigar a los pobres*, 2.ª ed., trad. de Gemma Deza. Madrid: Capitan Swing.

Ferguson, A. G. (2017). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. New York University Press.

Floridi, Luciano (2022). *Etica dell'intelligenza artificiale* Sviluppi, opportunità, sfide. Cortina Raffaelo.

Friedler, Sorelle A; Scheidegger, Carlos & Venkatasubramanian, Suresh (2016). On the (im)possibility of fairness. https://arxiv.org/abs/1609.07236

Goldman, Barry & Cropanzano, Russell (2015). "Justice" and "fairness" are not the same thing", *Journal of Organizational Behavior*, 36, 313-318. DOI: 10.1002/job.1956

Hardt, Moritz; Price, Eric & Srebro, Nati (2016). Equality of opportunity in supervised learning. En: *Advances in Neural Information Processing Systems*.

Hobson, Zoë; Yesberg, Julia A.; Bradford, Ben & Jackson, Jonathan (2021). Artificial fairness? Trust in algorithmic police decision-making, *J Exp Crimino*l, 1-25.

Huang, Wenxuan (2022). Reduce model unfairness with maximal-correlation-based fairness optimization, Master Thesis.
https://repository.tudelft.nl/islandora/object/uuid%3A8f40561a-80be-4047-9760-63ab27207ffc

Katz, Yarden (2017). *Manufacturing an Artificial Intelligence Revolution: Neoliberalism and the 'new' big data Yarden Katz*. Harvard, Harvard University.

Katz, Yarden (2020). Artificial Whiteness. *Politics and Ideology in Artificial Intelligence.* Columbia University Press.

Kleinberg, Jon; Mullainathan, Sendhil & Raghavan, Manish (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of Innovations in Theoretical Computer Science (ITCS 2017).* https://arxiv.org/abs/1609.05807

Kleinberg, Jon; Ludwig, Jens; Mullainathan, Sendhil & Rambachan, Ashesh (2018). Advances in big data research in economics. Algorithmic fairness. *AEA Papers and Proceedings*, 108, 22-27.

Konstantinov, Nikola (2022). "Encontrar la equidad en la IA" (entrevista realizada por Sandrine Ceurstemont (17 de mayo de 2022).
https://topbigdata.es/encontrar-la-equidad-en-la-ia-noticias/

Kraus, Rachel (2018). Amazon used AI to promote diversity. Too bad it's plagued with gender bias. Algorithms reflect societal biases in more ways than one. October 10, 2018. https://mashable.com/article/amazon-sexist-recruiting-algorithm-gender-bias-ai

Kusner, Matt; Lofthus, Joshua; Russell, Chris & Silva Ricardo (2017). Counterfactual fairness, Neural Information Processing Systems. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Galeotti, Mattia (2018). Discriminazione e algoritmi. Incontri e scontri tra diverse idee di fairness, *The lab's quarterly*, XX(4) 73-96.

Llano Alonso, Fernando Higinio (2018). *Homo excelsior. Los límites jurídicos del transhumanismo,* Valencia: Tirant lo Blanch.

Lledó Yagüe, Francisco (2022). *Los nuevos esclavos digitales del siglo XXI y la superación del hombre óptimo. ¿Hacia un nuevo derecho robótico?* Madrid: Dykinson.

MacIntyre, Alasdair (1988) [2001] *Justicia y racionalidad: conceptos y contextos*. Trad. de Alejo Jose G. Sison, Editorial: Eiunsa.

Martínez García, Jesús Ignacio (2020). La respuesta jurídica. *Anuario de Filosofía del Derecho*, 36, 347-371.

Mehrabi, Ninnareh; Morstatter, Fred; Saxena, Nripsuta; Lerman, Kristina & Galstyan, Aram (2021). A survey on Bias and fairness in Machine Learning", arXiv:1908.09635v3 [cs.LG https://arxiv.org/abs/1908.09635

Miller, David (2021). Justice. En: Edward N. Zalta (editor), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Nussbaum, Martha (2006). *Las fronteras de la justicia. Consideraciones sobre la exclusión.* Barcelona: Paidós.

O'Neil, Kate (2016). *Weapons of math destruction. How big data increases inequality and threatens democracy*, Penguin, New York, 2016 [trad. Española (2018): *Armas de destrucción matemática: como el big data aumenta la desigualdad y amenaza la democracia*, trad. de Violeta Arranz de la Torre, Capitán Swing].

Perez-Luño, Antonio Enrique (1996). *Manual de Informática y Derecho.* Barcelona: Ariel.

Rafanelli, Lucía M. (2022). Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy, *Big Data & Society,* January-June: 1-5. https://journals.sagepub.com/doi/pdf/10.1177/20539517221080676

Rawls, John [1971] (2006). *Teoría de la Justicia*, trad. de María Dolores González, 6.ª reimpresión.

Rawls, John [2001] (2012). *La justicia como equidad. Una reformulación,* trad. Andrés de Francisco. Edición a cargo de Erin Kelly. Barcelona: Paidós/Estado y Sociedad.

Santangelo, Antonio (2020). Equità degli algoritmi e democrazia*, DigitCult*, 5(2), 21-30. http://dx.doi.org/10.53136/979125994120634

Sandel, Michael (2009). *Liberalismo y los límites de la justicia.* Barcelona: Gedisa.

Sandel, Michel (2018). *Justicia. ¿Hacemos lo que debemos?* Barcelona: Penguin Random House.

Saxena, Nripsuta Ani; Huang, Karen; DeFilippis, Evan; Radanovic; Goran; Parkes, David. C. & Liu, Yang (2019). How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness, pp. 1-8.
https://econcs.seas.harvard.edu/files/econcs/files/saxena_ai19.pdf

Scanlon, Thomas M. (1998). *What We Owe to Each Other*, Belknap Press of Harvard University Press.

Scanlon, Thomas M. (2020). *Why does inequality matter?* Oxford: Oxford University Press.

Sen, Amartya (1980). Equality of What? En: McMurrin S. Tanner *Lectures on Human Values*, Volume 1. Cambridge: Cambridge University Press.

Solar Cayón, Jesús Ignacio (2019). *La Inteligencia Artificial Jurídica. El impacto de la innovación tecnológica en la práctica del Derecho y el mercado de servicios jurídicos*, Cizur Menor (Navarra): Aranzadi.

Soriano, Alba (2021). Decisiones automatizadas y discriminación: aproximación y propuestas generales, *Revista General de Derecho Administrativo*, 56, 1-45.

Stewart, Matthew (2020). Cómo lograr la equidad en los algoritmos. https://www.codetd.com/es/article/12010244

Tang, Zeyu; Zhang, Jiji & Zhang, Kun. (2022). What-Is and How-To for Fairness in Machine Learning: A Survey, Reflection, and Perspective**,** arXiv preprint. arXiv:2010.04053

Verma, Sahil & Rubin, Julia (2018). Fairness Definitions Explained, 2018 ACM/IEEE *International Workshop on Software Fairness*, 1-7.

Walzer, Michael (2001). *Las esferas de la justicia: una defensa al pluralismo y la igualdad*. México: Fondo de Cultura Económica, 2.ª edición.

Wang, Zhao (2021). *Fairness-aware multi-task and meta learning*. Dissertation presented to the Faculty of the University of Texas at Dallas. Doctor of philosophy in computer science.

Wang, Zhao & Shu, Kai (2021). Enhancing Model Robustness and Fairness with Causality: A Regularization Approach. Conference: Proceedings of the First Workshop on Causal Inference and NLP, DOI:10.18653/v1/2021.cinlp-1.3

Xiang, Alice & Raji, Inioluwa Deborah (2019). On the Legal Compatibility of Fairness Definitions, arXiv preprint arXiv:1912.00761, 1-6.

Young, Iris Marion (2011). *Responsibility for Justice*, New York: Oxford University Press.

Young, Iris Marion (2000). *La justicia y la política de la diferencia*. Trad. de Silvina Álvarez. Madrid: Ediciones Cátedra.

Zafar, Muhammad Bilal; Valera, Isabel; Gómez Rodríguez, Manuel & Gummadi, Krishna P. (2017). Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. En *Proceedings of the 26th International Conference on World Wide Web*

Informes

Fairness e Machine Learning. Il concetto di equità e relative formalizzazioni nel campo dell'apprendimento automatico. Nexa Center for Internet & Society Working paper nr 2/2018, Politécnico de Torino. https://nexa.polito.it/nexacenterfiles/Articolo%20TIM.pdf

IBM Policy Lab. Cómo mitigar el sesgo en los sistemas de Inteligencia Artificial (8 de junio de 2021). https://www.ibm.com/blogs/policy/latin-america/2021/06/08/como-mitigar-el-sesgo-en-los-sistemas-de-inteligencia-artificial/

Nuevas normas sobre la Inteligencia Artificial: preguntas y respuestas. Comisión Europea (21 de abril de 2021). https://ec.europa.eu/commission/presscorner/detail/es/qanda_21_1683

Informe *Automating Society 2020*. Algorithm Watch. https://automatingsociety.algorithmwatch.org/

Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión, COM (2021) 206 final, 2021/01106 (COD), 21/4/2021. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex:52021PC0206