Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 63-80
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.562811

Cerrando una brecha: una reflexión multidisciplinar sobre la discriminación algorítmica*,**

Bridging a gap: A multidisciplinary reflection on algorithmic discrimination

*PILAR DELLUNDE****
*ORIOL PUJOL*****
*JORDI VITRIÀ******

Resumen. Este artículo aborda el concepto de discriminación algorítmica desde una perspectiva conjunta de la filosofía y la ciencia de la computación, con el propósito de establecer un marco de discusión común para avanzar en el despliegue de las inteligencias artificiales en las sociedades democráticas. Se presenta una definición no normativa de discriminación y se analiza y contextualiza el concepto de algoritmo usando un enfoque intencional, enmarcándolo en el proceso

Abstract. This article presents a joint reflection from philosophy and computer science on the concepts behind algorithmic discrimination with the aim of providing a common framework for discussion to advance the deployment of artificial intelligence in democratic societies. A non-normative definition of discrimination is presented, and the concept of algorithm is analyzed and contextualized using an intentional approach,

Recibido: 28-03-2023. Aceptado: 27-06-2023

* Esta publicación ha sido parcialmente financiada por los proyectos: 2021 SGR 01104 y 2021 SGR 00754 de la Generalitat de Catalunya y H2020-MSCA-RISE-2020 project MOSAIC (Grant Agreement 101007627).

** Todos los autores han participado de la misma forma en todo el proceso de elaboración del trabajo y han puesto sus nombres en orden alfabético de apellidos.

*** Es Catedrática de Lógica del Departament de Filosofia de la Universitat Autònoma de Barcelona. Sus líneas de investigación se centran en la lógica fuzzy, argumentación computacional y ética en el diseño de sistemas de inteligencia artificial. Ha publicado recientemente "An art painting style explainable classifier grounded on logical and commonsense reasoning" (Soft Computing, 2023) y "Probabilistic Argumentation: An Approach Based on Conditional Probability" (Lecture Notes in Computer Science 12678, 2021). Contacto: pilar.dellunde@uab.cat

**** Es Catedrático del Departament de Matemàtiques i Informàtica de la Universitat de Barcelona. Su línea de investigación principal trata sobre los fundamentos algorítmicos del aprendizaje automático y su impacto. Ha publicado recientemente "The Forgotten Human Autonomy in Machine Learning" (CEUR-WS, IAIL, 2022) y "Copying Machine Learning Classifiers" (IEEE Access 8, 160268-160284). Contacto: oriol_pujol@ub.edu

***** Es Catedrático del Departament de Matemàtiques i Informàtica de la Universitat de Barcelona. Sus líneas de investigación son aprendizaje automático, inferencia causal y aspectos éticos de la inteligencia artificial. Ha publicado recientemente "Estimand-Agnostic Causal Query Estimation with Deep Causal Graphs" (IEEE Access 10, 2022, 1370-71386), y "A survey on uncertainty estimation in deep learning classification systems from a Bayesian perspective" (ACM Computing Surveys, 54 (9), 2022, 1-35). Contacto: jordi.vitria@ub.edu

de toma de decisiones e identificando las fuentes de discriminación, así como los conceptos detrás de su cuantificación para terminar exponiendo algunos límites y desafíos.
Palabras clave: Discriminación, algoritmos, inteligencia artificial, métricas.

framing it in the decision-making process and identifying the sources of discrimination, as well as the concepts behind its quantification, ending by exposing some limits and challenges.
Keywords: Discrimination, algorithms, artificial intelligence, metrics.

1. Introducción

En los últimos capítulos de *Homo Deus*, Y. N. Harari reflexiona sobre el concepto de dataísmo (Harari, 2016, 821): «El dataísmo sostiene que el universo consiste en flujos de datos, y que el valor de cualquier fenómeno o entidad está determinado por su contribución al procesamiento de datos.» afirmando que «el trabajo de procesar los datos debe encomendarse a algoritmos electrónicos, cuya capacidad excede con mucho a la del cerebro humano.»

El debate al que Y. N. Harari contribuye tiene su origen, entre otros, en el artículo de D. Brooks publicado en *The New York Times*, en el que sostiene que existe una revolución de los datos:

> Ahora tenemos la capacidad de acumular enormes cantidades de datos. Esta capacidad lleva consigo un cierto presupuesto cultural —que todo lo mensurable debe ser medido; que los datos son lentes transparentes y fiables que nos permiten filtrar todo emocionalismo y toda ideología; que los datos nos ayudarán a hacer cosas significativas como predecir el futuro. (Brooks, 2013).

Muchas de las reflexiones filosóficas relevantes sobre el dataísmo prestan poca atención a lo que Y. N. Harari llama *algoritmos electrónicos,* o a la intervención humana en los procesos de decisión que utilizan dichos algoritmos. La trasparencia y fiabilidad de los datos a la que hace referencia D. Brooks y que recurrentemente encontramos en el imaginario popular, el discurso gubernamental, el márquetin mercantilista, e incluso, en ocasiones, en el argumentario filosófico, implica diversas nociones que promueven imágenes de los algoritmos que los presentan como objetivos y de propósito universal. Por ejemplo, el panel de la Unión Europea, en su estudio "A governance framework for algorithmic accountability and transparency", indica

> As with much of the digital economy, the use of algorithmic systems is characterised by the highly cross-border nature and global reach of the services that are built on these technologies. (EU Res Service, 2019)

o en las críticas a esta supuesta objetividad en (Guersenzvaig et al., 2022). El discurso sobre la objetividad tiene como finalidad última legitimar la confianza en el algoritmo bajo la afirmación de "un algoritmo objetivo es confiable". El concepto de objetividad científica entendido como fidelidad a los hechos/datos, ideal libre de valores, e independiente de sesgos es incompatible con el del algoritmo basado en IA. Si bien es cierto que el algoritmo

tiene un comportamiento determinista y por lo tanto libre del sesgo intrapersonal que se encuentra en la decisión humana; es decir, dado un mismo estímulo/situación un ser humano puede decidir sobre este de forma diametralmente opuesta en función de su estado mental y de su interocepción. Pero, salvo este comportamiento determinista, los algoritmos basados en datos no están libre de valores y sesgos.

Como se comentará, el tecnólogo diseña un producto con una intencionalidad y propósitos concretos, para conseguir unos objetivos, usualmente de forma conjunta con los especialistas del dominio y los agentes que pondrán el sistema en funcionamiento. Fuera de posibles pretensiones por parte de grandes corporaciones tecnológicas, raramente un tecnólogo puede pretender por su cuenta crear un sistema con vocación universal. La consciencia de la necesidad de acotar el uso del algoritmo a un contexto y entorno determinado viene marcada de salida por la propia accesibilidad a los datos, por su representatividad, por el conocimiento de los potenciales sesgos estructurales que estos pueden tener y por los desbalanceos potenciales correspondientes a sesgos estadísticos en el propio diseño experimental y la recogida de datos. Todos estos factores hacen que en la literatura técnica de aprendizaje automático se hable de conceptos que precisamente reconocen la limitación de uso y de objetivos y estudian formas de combatir estas limitaciones. Por ejemplo, conceptos como *transfer learning, domain adaptation* (Pratt, 1993), o *differential replication* (Unceta, 2020), son conceptos propios de las disciplinas de aprendizaje automático que asumen que el mundo real es variable y está en constante cambio y que la representación que el algoritmo tiene de éste es particular y puede contener sesgos no deseados. Los conceptos anteriormente mencionados recogen formas de adaptar y reaprovechar el conocimiento aprendido en el algoritmo cuando el dominio o las restricciones que imperan sobre el algoritmo ya sean normativas, operacionales, éticas, etc., convierten al algoritmo en no usable.

Las contribuciones originales de este artículo son producto de una reflexión conjunta desde las ciencias de la computación y la filosofía sobre la IA, centrado en la discriminación algorítmica, con el objetivo de avanzar de forma efectiva en un diseño más ético de estas tecnologías. Más concretamente: en la sección 2, revisamos el concepto de discriminación desde un punto de vista filosófico, proponiendo una el uso de una definición no normativa del concepto de discriminación, que permite aislarla de conceptos normativos o moralizantes y que puede fácilmente ser compartida y resultar útil en el campo de las ciencias de la computación. En la sección 3 proponemos un marco de análisis de la discriminación en la toma de decisiones que va más allá del "algoritmo", basado en el enfoque intencional, que define un marco teórico que también puede ser compartido entre la IA y la filosofía. Esto nos permite introducir la idea de "reducción al enfoque de diseño" como criterio para separar problemas de discriminación simples (originados por una discrepancia en los criterios normativos en un sistema deductivo) de problemas complejos (en los que los criterios normativos son implícitos y mediados por los datos y el proceso de construcción del sistema inductivo). Posteriormente, analizamos en detalle los aspectos propios de la toma de decisiones basada en IA e identificamos las fuentes de la discriminación, introduciendo el concepto de *legitimidad del uso de un sistema de IA*, como estadio previo a cualquier análisis ético del uso de la IA. En la sección 4, ampliamos la visión del *algoritmo como sistema* a *algoritmo dentro del proceso de la toma de decisiones*, e introducimos los conceptos fundamentales

y líneas de pensamiento filosófico que sirven de guía a la cuantificación de la discriminación. Así mismo, subrayamos las dificultades en este proceso e identificamos vías para uso efectivo. Finalmente, en la última sección exponemos algunos de los límites actuales y la necesidad de una reflexión multidisciplinar que genere un marco de discusión común para poder avanzar en el despliegue y uso de la IA en una sociedad democrática, teniendo en cuenta el punto de vista de todos los agentes implicados.

2. Una aproximación filosófica a la discriminación

En el artículo "Bias and Fairness in Machine Learning" (Mehrabi et al., 2021) de la serie *ACM Computing Surveys*, los autores presentan un estado del arte sobre los sesgos y la equidad en el aprendizaje automático, e introducen una taxonomía para clasificar las diferentes definiciones de equidad con las que los investigadores en aprendizaje automático han trabajado para tratar de evitar los sesgos en los sistemas de IA. En la parte final del artículo, se discuten retos y oportunidades para futuras investigaciones en este campo. Llama la atención el primero de estos retos que ellos definen como *synthesizing a definition of fairness*. Este desafío plantea varios elementos interesantes para la reflexión filosófica. En primer lugar, al plantear este reto no se tiene en cuenta que los conceptos de equidad y discriminación son complejos y dinámicos, y no pueden ser representados únicamente utilizando formalismos matemáticos, ya que su significado está condicionado por el contexto y el momento histórico.

En segundo lugar, aunque algunos autores, por ejemplo, (Simon et al., 2020) y (Seng et al., 2021), han destacado la necesidad de adoptar un enfoque más holístico en el diseño de estas tecnologías, cabe destacar que esta sigue siendo una demanda común (incluso en proyectos interdisciplinarios que emplean metodologías como el *Value Sensitive Design*), el pedir a los investigadores de humanidades y ciencias sociales que proporcionen definiciones que sirvan como base para introducir una noción precisa, universal y formal para su implementación. Esto implica desconocimiento del papel de las definiciones en estos ámbitos, y plantea la necesidad de resituar el papel de las humanidades y las ciencias sociales en el diseño de los sistemas de IA.

En esta sección presentamos la definición filosófica de discriminación del libro (Lippert-Rasmussen, 2014), una definición no moralizante (que no implica necesariamente que discriminar es incorrecto) llamada *discriminación grupal*. Esta definición representa un concepto más amplio de discriminación que el que encontramos en los artículos sobre discriminación algorítmica. Si bien en algunos artículos sobre discriminación algorítmica se citan contribuciones de K. Lippert-Rasmussen, por ejemplo, respecto a la discriminación estadística (Barocas, 2016), las reflexiones de (Lippert-Rasmussen, 2014) son aún poco conocidas en el ámbito de las ciencias de la computación. Sin pretender hacer una presentación exhaustiva, introducimos aquellos elementos básicos de la definición que consideramos que pueden ser útiles para futuras reflexiones interdisciplinares en este ámbito.

En (Lippert-Rasmussen, 2014) se presentan condiciones necesarias y suficientes para considerar que un acto, una política o una práctica es discriminatoria a nivel de grupo:

> X discriminates against Y in relation to Z by Φ-ing if, and only if,
>
> (i) there is a property, P, such that (X believes that) Y has P and (X believes that) Z does not have P,
>
> (ii) X treats Y worse than Z by Φ-ing,
>
> (iii) it is because (X believes that) Y has P and (X believes that) Z does not have P, that X treats Y worse than Z by Φ-ing,
>
> (iv) P is the property of being member of a certain socially salient group (to which Z does not belong), and
>
> (v) Φ-ing is a relevant type of act etc., and there are many acts etc. of this type, and this fact makes people with P (or some subgroup of these people) worse off relative to others, OR Φ-ing is a relevant type of act etc., and many acts etc. of this type would make people with P worse off relative to others, OR X's Φ-ing is motivated by animosity towards individuals with P or by the belief that individuals who have P are inferior or ought not to intermingle with others. (Lippert-Rasmussen, 2014, 44-45)

Hemos elegido esta definición porque no es moralizante, y precisamente por ello, nos permite analizar casos como el de los impuestos proporcionales, que tratan de manera diferente a los contribuyentes; la aplicación de este criterio proporcional representa un trato desventajoso para las personas con más capacidad económica, pero podemos considerar que esta discriminación no es ni incorrecta ni injusta. Una característica importante de la definición es que discriminar implica un trato desventajoso, hecho que debe distinguirse del trato que causa daños.

La creación de un marco conceptual para el debate sobre la discriminación algorítmica hace que un concepto que tenga en cuenta la pertenencia a un grupo sea más útil que una concepción como la de B. Eidelson que sostiene que «acts of discrimination are intrinsically wrong when and because they manifest a failure to show the discriminatees the respect that is due them as persons.» (Eidelson 2015: 7) Este relato prescinde explícitamente del requisito de pertenecer a un grupo socialmente destacado, y en su lugar solo requiere que el discriminador responda a alguna diferencia percibida de cualquier tipo entre la víctima y otras personas. Eidelson considera dos dimensiones de la personalidad (*personhood*): todas las personas 1) son iguales y tienen valor intrínseco y 2) son agentes autónomos. (Eidelson, 2015: 79), y la discriminación puede violar una o ambas de estas dimensiones. Este enfoque implica utilizar una definición moralizante de discriminación.

Para K. Lippert-Rasmussen es central la relación que se establece entre discriminación y trato diferencial en base a la pertenencia a un grupo socialmente destacado (*socially salient group)*. El autor entiende que un grupo es socialmente destacado si la percepción de la pertenencia a él es importante en la estructura de las interacciones sociales a través de una amplia gama de contextos. Ejemplos de grupos socialmente destacados pueden ser el conjunto de las mujeres o el de las personas inmigrantes en un país en un momento determinado de tiempo. En cambio, esta definición grupal nos permite no considerar, de manera general, la meritocracia como una discriminación de grupo.

La discriminación es esencialmente comparativa, ya que no se puede discriminar a nadie a menos que haya otras personas que reciban un trato mejor en comparación. Otra importante característica de la discriminación así definida es que es independiente de las propiedades reales de las personas, ya que no hay necesaria superposición entre las propiedades que convierten a alguien en objeto de discriminación y las propiedades que la persona realmente posee. Por ejemplo, un hombre podría ser víctima de discriminación contra las mujeres, por un error en la entrada de sus datos.

De especial relevancia para el estudio de las implicaciones éticas de los sistemas de toma de decisiones es el concepto de discriminación estadística, es decir, cuando se trata a personas de manera diferente sobre la base de generalizaciones estadísticas explícitas o implícitas sobre el grupo al que esta persona pertenece. En tanto que caso particular de la definición de discriminación que hemos introducido al principio de la sección, la discriminación estadística es esencialmente comparativa y se discrimina en función de la pertenencia a un grupo socialmente destacado, a excepción de casos como la discriminación genética. Lippert-Rasmussen añade una cláusula más a la definición original para definir este tipo de discriminación:

> (vi) It is because (X believes that) Y has P and (X believes that) Z has not, and because (X believes that) P is statistically relevant, that X treats Y worse than Z by Φ-ing. (Lippert-Rasmussen 2014: 81)

Pero ¿qué significa que un grupo socialmente destacado sea estadísticamente relevante? En (Lippert-Rasmussen 2014: 86) se considera que un grupo es estadísticamente relevante si la probabilidad de tener alguna otra característica (por ejemplo, solicitar permiso de paternidad, o poseer drogas ilegales) varía sobre la base de qué grupos uno es miembro.

A veces, la evidencia estadística disponible puede ser correcta y utilizada de una manera no sesgada. La discriminación estadística no necesariamente se basa en evidencias estadísticas insuficientes o falsas. Si bien el uso de información estadística a menudo puede ser selectivo (específicamente, a menudo se puede usar para tomar como objetivo a minorías), el uso de la información estadística *per se* no necesita ser selectivo. No todo tipo de discriminación estadística está relacionada con creencias sobre un estatus inferior de aquellos a quienes se discrimina. Por ejemplo, la discriminación estadística a nivel tributario tiene como objetivo evitar la evasión de impuestos por parte de los más ricos.

En esta sección hemos dejado fuera elementos importantes, como el análisis en profundidad de la definición de *socially salient group*, de discriminación indirecta (muy relevante porque es una de las más difíciles de detectar en el diseño de los sistemas de IA) y que el propio (Lippert-Rasmussen 2014: 54-74) construye sobre la definición aquí presentada, de discriminación económica, o de interseccionalidad. Nuestro objetivo no era una presentación exhaustiva, sino, sobre todo, hacer explícita la complejidad del debate sobre esta noción, y cuestionar la posibilidad de encontrar una única formalización operacional de este concepto.

3. Sobre los algoritmos y la discriminación

En el imaginario colectivo sobre la discriminación algorítmica existen varios conceptos (algoritmo, datos, sesgos, etc.) que predefinen las bases de la discusión y que desde nuestro punto de vista requieren de una mayor elaboración para convertirse en fundamentos sólidos de un diálogo fructífero entre la filosofía y las ciencias de la computación. En los siguientes apartados se hace un análisis crítico de su significado y se propone pasar del concepto de *algoritmo* al concepto de *sistema*.

3.1. Algoritmos, sistemas de inteligencia artificial y el enfoque intencional

La idea de discriminación se produce a un nivel de abstracción que requiere el concepto de agencia y desde este punto de vista creemos que el uso del término "algoritmo" no es especialmente adecuado, a causa de su significado reduccionista.

El concepto clásico de "algoritmo", entendido como una secuencia de instrucciones que sirven para llegar a solución de un problema, evoca un enfoque deductivo que no representa los actuales sistemas de inteligencia artificial (IA). Los sistemas de IA se sitúan en una categoría especial de sistemas que, siguiendo el marco teórico de D. Dennett (Dennett, 1987), deben ser entendidos desde un enfoque intencional y que tienen una naturaleza distinta de los algoritmos clásicos.

Según la propuesta de D. Dennett tenemos tres alternativas cuando queremos entender un sistema complejo. La primera alternativa, el enfoque físico, usa las leyes de la física a un determinado nivel de abstracción para modelizar el sistema a partir de sus constituyentes y de las interacciones que podemos observar. El comportamiento de un paraguas que sale volando a causa del viento estaría a este nivel. Desde este punto de vista "entender" el sistema quiere decir tener una cierta capacidad de predicción de su comportamiento usando exclusivamente las leyes fundamentales de la naturaleza.

La segunda alternativa, el enfoque de diseño, nos permite entender un sistema a partir de la asunción que ha sido diseñado con un propósito y que por lo tanto cabe esperar que su comportamiento se ajuste a este propósito. En este caso los aspectos ligados a la física del sistema pueden ser subsidiarios y por lo tanto no especialmente útiles para entender su comportamiento. Una silla es un claro ejemplo de sistema que debe ser interpretado con un enfoque de diseño, al igual que un reloj o un procesador de textos en un ordenador.

La tercera alternativa, o enfoque intencional, se aplica a aquellos sistemas que se entienden mejor como agentes racionales, a los que se puede suponer unas creencias, un propósito y hasta una cierta representación del mundo que les permite conseguir su propósito. Pertenecen a este nivel sistemas tan dispares como un termostato, una colonia de hormigas, un ser humano, una empresa o la misma sociedad en la que vivimos, pasando por sistemas artificiales complejos, como robots o sistemas de IA.

Situados en este marco de análisis, podríamos preguntarnos si las acciones de un termostato son potencialmente discriminatorias para aquellas personas con una sensación térmica fuera de los rangos que el termostato supone normales. Desde el punto de vista intencional el termostato tiene un propósito y unas creencias bien definidos, así como un comportamiento

que le permiten conseguir su propósito en la mayoría de los casos. Suponiendo pues este escenario, el análisis de este tipo de discriminación desde el enfoque intencional no tiene mucho recorrido porque sus propósitos, creencias y comportamiento se pueden traducir sin ambigüedades a una serie de decisiones de diseño. Esto permite reducir el problema de la discriminación, en el caso del termostato, a un problema de diseño y buscar una solución sin movernos de ese nivel de abstracción, en el cual los humanos nos sentimos especialmente cómodos desde hace siglos.

El caso de un algoritmo desarrollado por un programador y que tiene por objetivo codificar una serie de reglas usadas, por ejemplo, en un proceso burocrático de concesión de ayudas sociales (Johnson, 2022), se encuentra exactamente al mismo nivel que el termostato y también puede ser reducido a un problema de diseño, puesto que las cuestiones normativas que se pueden derivar de su definición y uso se reducen a dos escenarios: una discrepancia en el valor normativo de las reglas o la presencia de un error en la codificación de las mismas.

La reducción del problema de discriminación a un problema de diseño también se extiende a sistemas complejos como un automóvil o un avión, pero no a un sistema de IA basado en datos. En este último caso, la reducción a nivel de diseño no es posible a causa de la naturaleza de su proceso de creación y de sus operaciones (Zerilli, 2022), mucho más complejo que los programas y algoritmos convencionales.

En los siguientes subapartados repasaremos los elementos del proceso de creación, desarrollo y despliegue de un sistema de IA y su papel en la perpetuación o amplificación de la inequidad. Siguiendo con la nomenclatura clásica en temas de discriminación algorítmica, usaremos el término *sesgo* para denotar alguna causa potencial de discriminación (Fazelpour, 2020), aun cuando el término es ambiguo y puede usarse con connotaciones positivas o negativas a constructos tan distintos como un algoritmo, un conjunto de datos o un comportamiento.

3.2. Sistemas de IA basados en datos

El elemento principal de cualquier teoría de la (in)equidad de la IA es que los modelos de IA basados en datos construyen un sistema predictivo para la toma de decisiones de forma inductiva, a diferencia de los métodos clásicos, que se basan en un proceso deductivo. Es esta diferencia la que determina, en la mayoría de los casos, el uso de los enfoques de diseño o intencional para su comprensión. El proceso inductivo es complejo e involucra una gran variedad de elementos de naturaleza distinta, hecho que refuerza el concepto de *sistema* en substitución de la de *algoritmo*. Por otra parte, este proceso puede ser origen de sesgos que resulten en problemas de discriminación.

3.2.1. Legitimidad

El proceso de creación de estos sistemas se inicia a partir de una descripción genérica de un objetivo que, en el caso de predecir eventos o estados futuros sobre personas y tener consecuencias sobre el mundo real, debe estar sujeto a un análisis ético. Suponiendo que este objetivo es legítimo (Sternberger, 1968), la legitimidad del uso de un sistema de IA para con-

seguir tal objetivo se puede determinar a partir del nivel de precisión de sus predicciones, de sus potenciales efectos discriminatorios, de su eficacia respecto al objetivo y también a veces del nivel de transparencia (Lazar, 2022). A estas propiedades podríamos añadir una condición de prudencia, la irreductibilidad: que no exista una solución viable basada en algoritmos clásicos y que por tanto el problema de la discriminación no sea reducible al enfoque de diseño.

3.2.2. Los datos

La adquisición de datos para entrenar un algoritmo de IA es un proceso que requiere un análisis detallado que permita evaluar su representatividad, veracidad, estabilidad, etc. y así evitar o minimizar los sesgos de representación (Mehrabi, 2021) o de medida (Jacobs, 2021).

Un aspecto de especial importancia en el proceso de recogida de datos es que el mundo real, *el mundo tal y como es*, no se corresponde necesariamente con nuestras creencias, con *el mundo tal y como podría y debería ser*. Los datos que obtenemos pueden ser pues reflejo de situaciones o procesos que pueden ser éticamente deficientes desde nuestro sistema de valores y que por lo tanto pueden contaminar el sistema resultante. Este tipo de sesgo se llama sesgo estructural, sistémico o social (Mitchell, 2021).

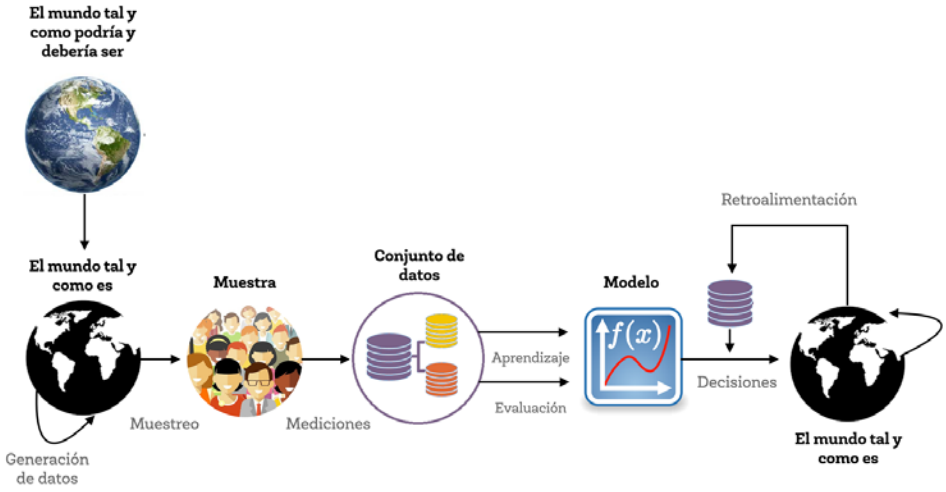

Figura 1. Proceso de creación de un sistema de IA

3.2.3. Los algoritmos

Dado un conjunto de datos que representa el fenómeno que se quiere modelizar, hace falta determinar una función objetivo, que pueda ser optimizada por un algoritmo de aprendizaje automático (Hardt, 2022). Esta función objetivo suele estar relacionada sólo de forma

indirecta con el objetivo general del sistema, puesto que éste no suele ser fácilmente expresable en términos predictivos. Por este motivo, si el objetivo general y la función objetivo no están alineadas se pueden introducir sesgos discriminatorios durante la predicción.

El producto final del algoritmo de aprendizaje es una serie de parámetros que definen un modelo de predicción. Estos modelos varían en su capacidad expresiva y también en su interpretabilidad, hecho que debe ser tenido en cuenta en aquellos ámbitos en los que la transparencia es necesaria (Rudin, 2019). También es importante señalar que en general los algoritmos de aprendizaje forman parte de librerías externas que han sido desarrolladas con objetivos genéricos y que su inspección no siempre está al alcance de los desarrolladores.

El algoritmo de aprendizaje también puede introducir otro tipo de sesgos que, aunque sean más sutiles que los originados en los datos, también pueden tener consecuencias. Este es el caso del sesgo inductivo (Mitchell, 1980) de un algoritmo de aprendizaje, entendido como el conjunto de suposiciones que se usan para maximizar la capacidad de generalización de un modelo.

3.2.4. El despliegue

El despliegue y aplicación de un modelo predictivo es uno de los aspectos menos estudiados, pero puede ser también el origen de alguna discriminación.

En primer lugar, se puede dar el llamado sesgo de despliegue, que se refiere a cualquier sesgo que surja cuando un sistema se usa o interpreta de manera inapropiada, muchas veces sin el conocimiento expreso de los diseñadores o desarrolladores. Este sesgo se podría dar, por ejemplo, en el caso de un modelo desarrollado para evaluar la probabilidad de cometer un nuevo delito y que fuera reusado para tomar decisiones sobre la duración de la sentencia de un preso.

También relacionado con el despliegue, podemos tener un sesgo de retroalimentación. Este tipo de sesgo se produce cuando un modelo en sí mismo influye, a través de su efecto en el mundo real, en la generación de datos que se utilizan para entrenarlo. Los algoritmos de recomendación son especialmente susceptibles de estar afectados por este sesgo (Baeza-Yates, 2018).

4. Sistemas algorítmicos en contexto

4.1. El proceso de decisión y el proceso de predicción

La consideración de algoritmo como sistema no es la única generalización que es necesario hacer cuando se habla de éstos. En la literatura técnica es común mezclar los conceptos de decisión y de predicción, que, aunque relacionados, deben considerarse aparte. El sistema algorítmico exclusivamente realiza predicciones. Como tal, define modelos que nos permiten establecer medidas de probabilidad, riesgo, o cuantificar la perdida en términos de la diferencia entre lo deseado y lo obtenido. Un proceso distinto es el de la toma de decisiones. En teoría de la decisión (Von Neumann, 1944) se plantea el estudio de cuál es la mejor decisión para un agente racional cuando se enfrenta a la selección de distintas opciones

bajo incertidumbre. La teoría de la decisión necesita establecer para cada posible acción la probabilidad de cada potencial respuesta, así como el valor asociado al par acción-respuesta. Y es este punto el que permite la unión entre la decisión y la predicción. El algoritmo de predicción puede alimentar las probabilidades de las respuestas. Sin embargo, queda libre un punto básico en esta teoría utilitarista, y es el valor asociado al par. Es decir, si tomo una decisión A y el resultado es B, ¿qué valor tiene este resultado? El valor del producto cartesiano entre acciones y resultados se recoge en una matriz denominada *matriz de decisión*. En la literatura de aprendizaje automático no se hace esta distinción entre predicción y decisión y pragmáticamente se considera que la decisión que se debe tomar corresponde a la que tiene mayor probabilidad. Y efectivamente esto es así cuando la matriz de decisión asigna la misma utilidad a todas las decisiones. En ocasiones se habla muy livianamente de estos conceptos introduciendo el coste de los errores de un sistema algorítmico cuando se introduce la *matriz de confusión*. Esta matriz cuantifica la cantidad de predicciones en función del valor esperado e introduce los conceptos de verdaderos y falsos, positivos y negativos. Recordemos que los falsos positivos (FP) y los falsos negativos (FN) corresponden a los errores de predicción realizados. Asociados a esta matriz aparecen métricas parciales de rendimiento que fundamentaran las métricas estadísticas de medida de la discriminación.

En este punto nos gustaría ampliar el discurso que considera que el proceso de decisión se compone exclusivamente de un algoritmo monolítico. Supongamos un algoritmo entrenado para la predicción de una enfermedad infecciosa grave. Un sistema que minimiza la cantidad de errores cometidos implícitamente asume que el valor utilitario de los FP y de los FN es el mismo. Sin embargo, el contexto nos indica que, en este caso, no deberían tener el mismo valor. Un FN significa que el sistema predice 'no enfermedad' a un paciente que sí padece la enfermedad. El riesgo que comporta tomar una decisión basada en este sistema es muy alto. Por lo tanto, sería deseable que el número de FN sea el menor posible, idealmente cero, aunque eso implique aumentar el número de FP. Con esto evitaríamos que pacientes con la enfermedad quedasen sin tratar, con el riesgo epidemiológico que comportaría. Pero ¿hasta qué punto es admisible aumentar los FP? Es admisible diagnosticar una enfermedad a un paciente sano, tanto en cuanto el sistema de diagnóstico y de decisión no termina ahí. Pruebas complementarias o nuevos sistemas de predicción usados a continuación, pueden focalizarse en esta población para poder distinguir los casos de forma más fina. La visión monolítica del algoritmo independiente del proceso de decisión nos oculta esta visión que debe considerarse en el uso de estos sistemas.

4.2. El algoritmo me discrimina

¿Cómo sé que estoy siendo discriminado? ¿Se puede medir la discriminación algorítmica? ¿Tiene sentido hablar de algoritmos que discriminan? Y si es así ¿qué consideraciones se han de tener en cuenta? ¿Qué aproximaciones a éste?

Las anteriores preguntas, que aplican a cualquier agente de toma de decisiones, ponen de manifiesto la necesidad de un instrumento que permita identificar esta percepción. En un intento de formalizar la discriminación, la comunidad científica que estudia la justicia algorítmica ha propuesto múltiples métricas para poder identificar, medir y así ayudar a mitigar

los efectos de la disparidad bajo distintas perspectivas (Verma, 2018). Estas métricas se pueden dividir en tres grandes bloques en función de la corriente filosófica que siguen: desde el punto de vista del igualitarismo encontramos un conjunto de métricas estadísticas. Siguiendo un enfoque de justicia individual en la línea aristoteliana (Libro V de la *Ética nicomáquea*) de los "casos similares" se establecen un conjunto de métricas individuales basadas en el concepto de similitud. Finalmente, subyacente al discurso sobre el trato e impacto dispar se encuentra la identificación de las causas que dan pie a las métricas causales.

4.2.1. Medidas estadísticas

Desde un punto de vista del igualitarismo encontramos un conjunto de métricas estadísticas. En estas definiciones, tal y como se promueve en la corriente igualitaria, se pretende ecualizar las oportunidades indistintamente del colectivo y de los elementos no controlables o elegibles de los individuos. En este punto entra el concepto de identidad grupal como elemento no controlable y por el que dos individuos que pertenecen a dos colectivos distintos deben tener la misma consideración.

Para definir las métricas vamos a considerar tres elementos: llamaremos T al valor real esperado, O al valor de la predicción y P a la propiedad que define el grupo destacado.

Para motivar las medidas consideremos el caso de un proceso de decisión algorítmico que se usa para contratar potenciales candidatos a un trabajo. Imaginemos que tenemos sospechas de que el proceso de decisión puede tener un sesgo potencialmente discriminatorio en función del sexo. Para identificar una potencial discriminación podríamos preguntarnos las siguientes preguntas: *¿Se contratan de la misma forma candidatos del sexo masculino y del sexo femenino? ¿Del conjunto de los candidatos que el algoritmo ha predicho que podían ser potencialmente contratados, el algoritmo hace diferencias entre los dos colectivos?* o *¿Sobre todos aquellos que merecen ser contratados, la tasa de contratación es la misma independientemente del colectivo al que pertenecen?* Las tres preguntas son pertinentes y dan lugar a métricas diferentes. La primera, conocida como *independencia*, hace referencia a si la predicción es independiente del atributo sensible, y se puede formalizar como sigue

$$p(O=1 \mid P=a) = p(O=1 \mid P=b).$$

Usualmente se usa en relación a la decisión ventajosa y se la denomina *paridad demográfica o estadística*, y se encuentra asociada a la doctrina de discriminación directa o trato dispar.

La segunda, conocida como *suficiencia*, está relacionada con el concepto de precisión de la matriz de confusión y se formaliza

$$p(T=1 \mid O=1, P=a) = p(T=1 \mid O=1, P=b).$$

Así formalizada se la denomina *paridad predictiva* y si se exige para las decisiones positivas y negativas se denomina *igualdad de exactitud de uso condicional*.

Finalmente, la tercera, conocida como *separabilidad*, se relaciona con la sensibilidad, y se formaliza,

$$p(O=1|\ T=1, P=a) = p(O=1|\ T=1, P=b),$$

A esta formalización se la conoce como *igualdad de oportunidades*. Al igual que en los anteriores casos si se exige para las dos decisiones se obtiene la *igualdad de posibilidades* (equalised odds). Estas dos últimas se asocian a las formas de discriminación indirecta o impacto dispar.

Existen diversos teoremas que nos demuestran el delicado equilibrio entre estas tres definiciones, que son mutuamente excluyentes (Chouldechova, 2016). Esto significa que las tres medidas no se pueden cumplir a la vez. Por lo que, bajo este conjunto de medidas, cualquier agente que realice una decisión y use una o más de estas medidas para justificar que no realiza un trato dispar se puede demostrar que ejerce un impacto dispar, y a la inversa. Esto nos indica que el uso del concepto discriminación grupal en genérico es limitante y no se puede considerar como un problema de optimización matemática con solución única.

4.2.2. Medidas individuales

Por otro lado, encontramos las medidas basadas en la justicia individual y la visión aristotélica de *tratar los casos similares de forma similar* (Aristoteles, *Nicho*.). En otras palabras, argumenta que las personas deben ser tratadas según sus características y circunstancias individuales, en lugar de su identidad grupal. La identificación de esta medida se formaliza en las siguientes tres desigualdades:

A. $d(\text{'Y'},\text{'Z'})<\text{tolerancia}$

B. $p(O=1|I=\text{'Y'}) > p(O=0|I=\text{'Y'})$

C. $p(O=1|I=\text{'Z'}) < p(O=0|I=\text{'Z'})$

Dados los casos I='Y' y I='Z' correspondientes a dos individuos distintos, las desigualdades (B) y (C) indican que la decisión tomada para 'Y' y para 'Z' es distinta. La desigualdad (A) nos indica que la diferencia entre 'Y' y 'Z' es pequeña.

Este enfoque no está exento de dificultades puesto que se ha de definir la distancia entre los individuos representados por sus descriptores. Así como en los anteriores casos se enfatiza la noción de característica diferencial, en este caso nos plantea la cuestión de bajo que parámetros dos individuos son comparables. Por otro lado, la noción de justicia individual también tiene asociada la noción de grupo puesto que se compara el individuo con un colectivo definido por similitud (Binns, 2020). Y aún más, se produce un efecto de estandarización y de perdida de la noción de individualidad.

4.2.3. Causalidad y justicia contrafactual

Subyacente al discurso de disparidad se encuentra el concepto de causalidad. Sin embargo, los algoritmos descritos hasta el momento son algoritmos que explotan las correlaciones estadísticas entre los datos y la variable a predecir. Se mezclan causas y efectos. El estudio de la causalidad de una acción requiere de una intervención en el mundo real. Sin embargo, ésta puede ser irrealizable físicamente, tener un impacto inaceptable o no ético. Los modelos causales permiten estudiar el efecto causal a partir de datos observacionales bajo ciertas condiciones. Permiten responder preguntas causales a dos niveles: a nivel poblacional a partir de intervenciones y a nivel individual a partir de contrafactuales. Esto permite establecer nuevas métricas de igualdad en la denominada *justicia contrafactual* (Carey 2022). Formalmente, se extenderían las métricas anteriores bajo la perspectiva de las intervenciones. Por poner un ejemplo usando la notación causal, el concepto de independencia se puede extender al campo causal de la siguiente forma:

$$p(O=1 \mid Do(P=a)) = p(O=1 \mid Do(P= b)).$$

Donde el operador Do(P=a) indica la intervención que hace que un determinado atributo P tome el valor *a* para toda la población. Esta expresión se lee de la siguiente forma: Queremos que la probabilidad de la predicción ventajosa sea la misma cuando forzamos que la característica P tome el valor *a*, que cuando forzamos que la característica P tome el valor *b*. Esto respondería a ¿Cuál es el efecto causal sobre la proporción de contrataciones si la negociación la realiza un hombre o una mujer?. A nivel contrafactual se plantean preguntas más complejas y que incluyen los hechos y el contrafactual, como, por ejemplo, ¿se habría contratado a un candidato si su sexo fuese femenino (contra) sabiendo que no se ha contratado y su sexo es masculino (factual)? Y por lo tanto sería análogo causal a las medidas individuales.

4.2.4. Usando las métricas

El uso efectivo del concepto de discriminación requiere de la selección deliberada e intencional de una métrica particular en función del establecimiento de las prioridades correspondientes al contexto de aplicación. Si nos centramos en las medidas estadísticas, diversas guías (Ruf 2021) (Binns 2020) nos permiten identificar como usarlas. Por ejemplo, en aquellas ocasiones que el uso del sistema pretenda mitigar una desigualdad sistémica propiciando acciones que protejan a grupos menos privilegiados asumiendo que pretende revertir un sesgo estructural se entiende que se busca establecer políticas que obvian la causalidad. En este caso métricas basadas en independencia como la paridad demográfica sería recomendable. Si, por el contrario, consideramos la ausencia de sesgos estructurales deberíamos usar métricas derivadas de las de suficiencia o de separabilidad. O si consideramos las métricas de justicia individual, las métricas individuales no ajustadas velarían por las disparidades debidas a las decisiones personales, mientras que las ajustadas por el grupo ayudarían a mitigar sesgos estructurales. Usualmente la realidad nos plantea situacio-

nes complejas que acostumbran a mezclar ambas visiones y pueden requerir una selección amplia de métricas.

Complementaria a la visión estadística que sólo nos explicita la relación entrada-salida, la visión causal nos permite considerar las interacciones causales entre los distintos atributos. Esta visión es mucho más rica puesto que nos permite identificar explícitamente los mecanismos asociados a la discriminación directa e indirecta. Y, por lo tanto, medirlos. En este contexto se habla de discriminación de un camino causal si existe un efecto causal entre la variable protegida y el resultado siguiendo el camino que las une en el grafo causal. Usando el concepto de contrafactual asociado a los efectos directos, indirectos y espurios, (Plecko, 2023) establece una potencial reconciliación entre la independencia y la suficiencia, rompiendo efectivamente el teorema de imposibilidad. Para ello introduce el concepto de *necesidad de negocio*. Este concepto es un paralelo a la anterior intencionalidad del sistema y que permitía el uso de las métricas estadísticas. En este caso, la necesidad de negocio requiere identificar que atributos son atributos que deben obedecer a conceptos de impacto directo y cuales a impacto indirecto. Con esta identificación, se pueden medir las distintas magnitudes y establecer el equilibrio adecuado entre ellas.

5. Conclusiones

La necesidad de conectar una tecnología, cada vez más cercana a las actividades propias de los humanos, con los aspectos epistemológicos y normativos derivados de su desarrollo y uso requieren un punto de partida común entre diversas disciplinas, libre de *apriorismos* y simplificaciones estériles. En esta dirección hemos propuesto el uso de un concepto no normativo de discriminación que permite aclarar algunos aspectos epistemológicos. También hemos usado la imposibilidad de reducir los algoritmos de IA a un enfoque de diseño como su elemento definitorio principal, libre de tecnicismos innecesarios. Finalmente, hemos repasado el estado del arte en las métricas de discriminación, haciendo especial hincapié en los problemas asociados a su uso y sus potenciales soluciones. El presente texto pretende realizar la conexión de todos estos conceptos entendiendo que, como limitación de éste, un tratamiento exhaustivo de muchas de las ideas presentadas requiere de una discusión en mayor profundidad.

Aunque el nivel de la reflexión sobre la relación entre datos, algoritmos y decisiones ha avanzado mucho desde las propuestas dataístas de la década pasada, algunos de los desafíos identificados siguen sin una solución evidente, al tiempo que se crean nuevos problemas. A continuación, hacemos un repaso a algunos de los desafíos desde un punto de vista epistemológico y normativo.

5.1. Limites epistemológicos

El hecho que un sistema predictivo preciso no constituye necesariamente una buena base para la toma de decisiones es una evidencia científica que ha impactado en la investigación en IA de forma reciente y con resultados desiguales. Dada una aplicación, es necesario evaluar es si el objetivo propuesto es de naturaleza puramente predictiva o tiene alguna carac-

terística intervencional que impida una toma de decisiones basada en datos observacionales (Fernández-Loría, 2022). Las técnicas de inferencia causal representan la mejor aproximación a este problema (Pearl, 2018), pero su viabilidad a gran escala está aún por demostrar.

En algunas situaciones, la discriminación puede ser compuesta y aditiva, es decir, puede estar basada en muchos pequeños actos de discriminación que desembocan en una consecuencia grave al cabo del tiempo. Las métricas de discriminación no son capaces de detectar estas situaciones, que sólo pueden ser detectadas con métodos cualitativos (Narayanan, 2022).

Las métricas basadas en grupos, en general, tienden a ignorar los méritos de cada individuo en el grupo. Algunas personas pueden ser mejores para una tarea determinada que otras personas del mismo grupo, lo que no se refleja en las definiciones de equidad basadas en grupos (Mittelstadt, 2023). Este problema puede dar lugar a dos comportamientos problemático: (a) la profecía autocumplida en la que, al elegir deliberadamente a los miembros menos calificados del grupo protegido, colaboramos en la construcción de un mal historial para el grupo, y (b) el tokenismo inverso, donde al no elegir a un miembro bien calificado del grupo no protegido, uno de los objetivos del sistema se convierte en crear refutaciones convincentes para los miembros del grupo protegido que tampoco son seleccionados.

5.2. Problemas normativos

Desde el punto de vista normativo hace falta avanzar en un concepto de legitimidad para el uso de los algoritmos de IA en la toma de decisiones que considere, tal y como lo hace en el campo de la filosofía política, las condiciones de legitimización a la vez que las consecuencias de la renuncia a su uso (Martin, 2022).

Los avances tecnológicos han abierto de nuevo la definición de grupo protegido y la necesidad de considerar a los llamados grupos algorítmicos (Wachter, 2022). Estos son los grupos creados a partir de técnicas de perfilado algorítmico, que no se correlacionan con grupos legalmente protegidos. La opacidad en su uso por parte de las grandes compañías tecnológicas abre la posibilidad de una nueva fuente de discriminación oculta que hace falta clarificar.

Por último, hace falta reconsiderar la relación entre las características usadas por los modelos predictivos y su función (Creel, 2022). Estos modelos pueden usar tres tipos de características: características legítimas, características protegidas, y características arbitrarias. Supongamos un proceso de decisión para la concesión de un préstamo. El sueldo podría estar entre las primeras, el género entre las segundas, y el número de ascensores en el edificio en el que se halla el domicilio habitual podría considerarse arbitraria. ¿En qué casos es ético el uso de características arbitrarias? ¿En qué casos no se deberían usar nunca características arbitrarias?

Referencias

Aristoteles (1984), *Nichomachean Ethics,* Princeton University Press, Vol.3.1131a10–b15

Baeza-Yates, R. (2018), "Bias on the web". Commun. ACM. 61, pp. 54–61.

Barocas, S.; Selbst, A. D. (2016), "Big Data's Disparate Impact", Cal. Law Review, Vol.104

Binns R. (2020), "On the apparent conflict between individual and group fairness", ACM Proceedings of Int. Conf. on Fairness Accountability and Transparency in Machine Learning.

Brooks, D. (2013), "Opinion | The Philosophy of Data". New York Times, https:// www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html

Carey, A.; Wu, X. (2022), "The Causal Fairness Field Guide: Perspectives from social and formal sciences", Frontiers in Big Data, Vol 5.

Chouldechova, A. (2017), "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". Big data, 5(2), pp. 153-163.

Creel, K.; Hellman D. (2022), "The algorithmic Leviathan: arbitrariness, fairness, and opportunity in algorithmic decision-making systems." Canadian Journal of Philosophy 52.1. pp. 26-43.

Dennett, D. C. (1987), *The intentional stance*. MIT Press.

Eidelson, B. (2015), *Discrimination and Disrespect*. Oxford University Press.

EU P Serv (2019) "A Governance Framework for algorithmic accountability and transparency". Recuperado de: "https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2019)624262"

Fazelpour, S.; Danks, D. (2020), "Algorithmic bias: Senses, sources, solutions." Philosophy Compass 16.8: e12760.

Fernández-Loría, C.; Foster P. (2022), "Causal decision making and causal effect estimation are not the same… and why it matters." INFORMS Journal on Data Science 1.1, pp. 4-16.

Guersenzvaig, A.; Casacuberta, D. (2022), "La quimera de la objetividad algorítmica: dificultades del aprendizaje automático en el desarrollo de una noción no normativa de salud", IUES ET SCIENTIA, Vol 8 N 1, pp. 35-56.

Harari, Y. N. (2015), *Homo Deus: A Brief History of Tomorrow*. Random House. Traducción al castellano de la editorial Debate.

Hardt, M.; Recht, B. (2022), *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press.

Jacobs, A. Z.; Wallach, H. (2021), "Measurement and fairness". In Proceedings of the 2021 ACM Conference on fairness, accountability, and transparency, pp. 375-385.

Johnson, R. A.; Zhang, S. (2022) "What is the Bureaucratic Counterfactual? Categorical versus Algorithmic Prioritization in US Social Policy". In 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1671-1682.

Lazar, S. (2022), "Legitimacy, Authority, and the Political Value of Explanations". arXiv preprint arXiv:2208.08628.

Lippert-Rasmussen, K. (2014), *Born Free and Equal?*. Oxford University Press.

Martin, K.; Waldman, A. (2022), "Are algorithmic decisions legitimate? The effect of process and outcomes on perceptions of legitimacy of AI decisions". Journal of Business Ethics, pp. 1-18.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. (2021), "A Survey on Bias and Fairness in Machine Learning". ACM Comput. Surv. 54, 6, Article 115, pp. 1-35.

Mitchell, T. M. (1980), "The need for biases in learning generalizations ".New Jersey: Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ., pp. 184-191.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; Lum, K. (2021), "Algorithmic fairness: Choices, assumptions, and definitions." Annual Review of Statistics and Its Application, 8, pp. 141-163.

Mittelstadt, B. D.; Allo, P.; Taddeo, M.; Watcher, S.; Floridi, L. (2016), "The ethics of algorithms: mapping the debate". Big Data & Society, pp. 1-26.

Mittelstadt, B., Wachter, S., Russell, C.s (2023), "The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default". Available at SSRN: https://ssrn.com/abstract=4331652

Narayanan, A. (2022), "The limits of the quantitative approach to discrimination." James Baldwin lecture [transcript], Princeton University.

Pearl, J.; Mackenzie, D. (2018), *The book of why: the new science of cause and effect*. Basic books.

Plecko, D; Bareinboin, E. (2023) "Reconciling predictive and statistical parity: A causal approach", arXiv:2306.05059v1.

Pratt, L. Y. (1993), "Discriminability-based transfer between neural networks" (PDF). NIPS Conference: Advances in Neural Information Processing Systems 5. Morgan Kaufmann Publishers. pp. 204–211.

Rudin C. (2019), "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". Nat. Mach. Intell. 1(5), pp. 206–15.

Ruf, B.; Detyniecki, M. (2021), "Towards the Right Kind of Fairness in AI", arXiv:2102.08453v7.

Seng, M.; Floridi, L.; Singh, J. (2021), "Formalising tradeoffs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics". AI & Society, 1:529–544.

Simon, J.; Wong, P.; Rieder, G. (2020), "Algorithmic bias and the Value Sensitive Design approach". Internet Policy Review, 9(4):1-16.

Sternberger, D. (1968), "Legitimacy" in International Encyclopedia of the Social Sciences (ed. D.L. Sills) New York: Macmillan, Vol. 9, p. 244.

Unceta, I. (2020), "Environmental Adaptation and Differential Replication in Machine Learning", Entropy (Basel). 3;22(10):1122.

Verma, S.; Rubin, J. (2018), "Fairness definitions explained". IEEE/ACM Int Workshop on Software Fainess,

von Neumann, J.; Morgenstern, O. (1944), *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

Wachter, S. (2022), "The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law." arXiv preprint arXiv:2205.01166.

Zerilli, J. (2022), "Explaining Machine Learning Decisions". Philosophy of Science, 89(1), pp. 1-19. doi:10.1017/psa.2021.13