Facultad de Administración y Dirección de Empresas

BUSINESS ETHICS & ARTIFICIAL INTELLIGENCE

Clave: 201700439

MADRID | Mayo 2022

ÍNDICE:

1. INTRODUCCIÓN

Aunque la Inteligencia Artificial (IA) nace oficialmente en 1956, con la invención del programa de ordenador *Logic Theorist* por parte A. Newell y H. Simmon[7], las aplicaciones comerciales de la IA no llegan hasta la década de los 80. En esta época, las aplicaciones comerciales de la IA se enfocaban en resolver problemas contables, de procesos o de producción. Si nos trasladamos a la actualidad, vemos que la IA tiene ahora una versatilidad que está varios órdenes de magnitud por encima de su capacidad hace 40 años, con investigadores tratando de emular el funcionamiento del cerebro en máquinas o implementando sistemas capaces de analizar ingentes cantidades de datos que atañen a todos los aspectos de nuestras vidas.[8] En el mundo de la empresa, el foco fundamental de la aplicación de IA hoy en día es el tratamiento y análisis de datos masivos, con el fin de aumentar el rendimiento económico de la empresa.[9]

"Actualmente la ubicuidad e integración empresarial de los algoritmos hacen que estos influyan en nuestras vidas, a veces incluso de forma silenciosa. Pueden determinar si alguien es contratado, ascendido, si se le ofrece un préstamo, así como determinar qué anuncios políticos y noticias se muestran a cada individuo que usa un producto que tiene cómo parte integral la utilización de algoritmos. Sin embargo, todavía no existe un consenso académico acerca de quién, y en qué medida, tiene la responsabilidad ética derivada del uso de estos algoritmos en la esfera empresarial."[10]

El objetivo de este Trabajo de Fin de Grado es dar respuesta a cuestiones relacionadas con la responsabilidad de los desarrolladores respecto de sus algoritmos, la responsabilidad de las empresas que utilizan algoritmos con el fin de lucrarse, y el fundamento de la responsabilidad de cada *stakeholder* en el proceso de toma de decisiones con un componente algorítmico.

Los algoritmos son herramientas tecnológicas que, por su naturaleza, llevan integrada la presuposición de un conjunto particular de valores. Siguiendo a Martin,

---

[7] Benítez, R., Escudero, G., Kanaan, S., & Rodó, D. M. (2014). *Inteligencia artificial avanzada*. Editorial UOC. pp. 13.
[8] Ibídem.
[9] Ibídem.
[10] Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of business ethics*, *160*(4), 835-850.

podemos decir que los algoritmos son herramientas "cargadas de valores".[11] Los algoritmos no son herramientas neutras, dado que la aplicación de estos algoritmos, en la práctica, conlleva consecuencias morales, refuerza o socava principios éticos o permite que se coarten los derechos de las personas.[12]

Los algoritmos son un actor importante en numerosas decisiones éticas e influyen en la delegación de funciones y responsabilidades dentro de estas decisiones.[13] Esto es particularmente relevante en el plano de la responsabilidad en la toma de decisiones empresariales, tanto por las consecuencias que conllevan a nivel operativo, como por las consecuencias "imprevistas" que puedan surgir de su utilización.[14] Las empresas deben ser responsables no sólo de los valores imbuidos en los algoritmos que utilizan, sino también del diseño de la decisión algorítmica, determinando quién hace qué dentro de esta decisión.[15] Es la única forma en que podrá delimitarse la responsabilidad de los diferentes agentes de una decisión.[16]

El propósito de esta investigación es analizar cómo afecta el uso de algoritmos en el mundo empresarial a la responsabilidad moral de cada *stakeholders* involucrado en ese proceso, sea como empresa, desarrollador o usuario. Para concretar ese análisis, aplicaré lo investigado a un análisis de la responsabilidad de los *stakeholders* de Facebook.

2. MOTIVACIÓN

Actualmente, la mayoría de las empresas (en 2021, el 56%[17]) utilizan herramientas para el análisis de Big Data. Con la información extraída de esa ingente cantidad de datos, entrenan sus algoritmos, con el fin de que estos sean cada vez más precisos en su cometido. Cuando compramos algo en Amazon o vemos una película en Netflix, pensamos que es nuestra propia elección.[19] Probablemente, no es del todo así. los algoritmos influyen en un tercio de nuestras decisiones en Amazon y en más del 80% en

[11] Ibídem.
[12] Ibídem.
[13] Ibídem.
[14] Ibídem.
[15] Ibídem.
[16] Ibídem.
[17] McKinsey. (2021) *The State of AI in 2021*: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2021
[19] Columbus, L. (2017). *53% of companies are adopting Big Data analytics*. Recuperado el 5 de diciembre de 2021. Forbes: https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/

Netflix[20]. No solo es Martin quien habla de la "carga de valores" implícita en los algoritmos, sino que otros autores reputados como Kraemer, Kees van Overveld y Peterson apoyan esa opinión, con lo que, personalmente, estoy de acuerdo. Por ello, considero que es importante preguntarse e investigar quién tiene responsabilidad sobre los efectos derivados del uso de algoritmos en el plano empresarial, dado que no solo tiene consecuencias económicas, sino éticas y sociales.

Las compañías tecnológicas, en especial FANG (excluyendo a Apple), llevan más de 20 años influyendo a gran escala en el desarrollo de la opinión pública, de los gustos de los consumidores y teniendo efectos que, aún sin ser intencionados, pueden ser negativos para una parte considerable de la población.

Parte de esos efectos negativos nacen de una IA cuya función no es juzgar el tipo de impacto que tiene ese contenido (que promueve o deja de promover) para la sociedad, sino maximizar el beneficio de la empresa que lo utiliza. Debido al impacto que tienen estas empresas en la vida diaria de miles de millones de personas, realizar una investigación que trate de discernir quién, y en qué medida, tiene responsabilidad sobre los efectos de esos algoritmos es de vital importancia.

Este estudio podría servir para incrementar el cuerpo de conocimiento existente en el área, además de para orientar las políticas de decisión empresariales respecto al uso de algoritmos de Inteligencia Artificial.

3. OBJETIVOS

3.1 Objetivos generales

El objetivo general de este trabajo es investigar la relación que existe entre una de las herramientas con más potencial del mundo moderno (la IA, y en particular, los algoritmos) y uno de los aspectos más importantes en la vida de la persona, la vida de la empresa y el funcionamiento de la sociedad (la ética).

Tanto persona como empresa forman parte de la sociedad e influyen en su desarrollo. Dado el crecimiento en adopción de la IA por parte de las empresas, y teniendo

[20] Hosanagar, K. (2020). *A human's guide to machine intelligence: how algorithms are shaping our lives and how we can stay in control*. Penguin Books.

en cuenta su naturaleza exponencial, es de vital importancia analizar esta tecnología y sus implicaciones empresariales desde una perspectiva ética.

3.2 Objetivos específicos

Investigar en qué medida tienen responsabilidad los distintos *stakeholders*, incluyendo desarrolladores, directivos y usuarios, que utilizan algoritmos en el ámbito empresarial, sobre las consecuencias que se derivan de su uso. En el caso de Facebook, por ejemplo, el impacto que tiene en la salud mental de los niños, la división política que fomenta, o la personalización individualista del feed, entre otros aspectos.

Aplicar el mapa ético propuesto por Mittlestadt et al. para analizar la pregunta de investigación de forma exhaustiva, tratando de contribuir al aclaramiento de la responsabilidad de los distintos *stakeholders* de Facebook de forma práctica.

4. METODOLOGÍA

Para llevar a término la investigación utilizaré una metodología cualitativa. El procedimiento que va a llevarse a cabo para realizar la investigación es el siguiente: realizar un estudio del tema basado en una revisión de la literatura existente en el ámbito del uso de los algoritmos en la empresa, y en el ámbito del desarrollo y uso ético de algoritmos e Inteligencia Artificial, profundizando en lo respectivo a responsabilidad ética del desarrollador.

De esta primera investigación sacaré el marco teórico del trabajo, tratando de diseñarlo de forma inteligible para que cualquier persona que se disponga a leer el trabajo pueda entender su contenido e implicaciones, incluso si no es alguien versado en el ámbito de la Inteligencia Artificial.

A partir de ahí, entraré a investigar utilizando artículos académicos especialmente relevantes, enfocados en temas más concretos dentro de las áreas de ética, uso ético de la tecnología en la empresa, dilemas éticos en la industria del Big Data, entre otros. También me serviré de podcasts y libros sobre Inteligencia Artificial para desarrollar una perspectiva más holística sobre el tema. Para contextualizar el marco legal que atañe al

objeto del trabajo, llevaré a cabo la lectura de la legislación GDPR. Utilizaré el marco ético propuesto por Mittlestadt et al. para el análisis práctico de la responsabilidad de los distintos *stakeholders* de Facebook.

5. MARCO CONCEPTUAL

5.1 La tecnología de la Inteligencia Artificial: clasificación

De acuerdo con Towards Data Science, una web especializada en la divulgación de conocimientos sobre ciencia de datos e inteligencia artificial, la IA como tecnología se puede clasificar[21] de dos formas, ambas con distintos subtipos. Las dos formas consisten en clasificar según: 1) capacidades de la IA y 2) funcionamiento de la IA. Clasificando por capacidades, podemos distinguir entre tres tipos de IA: 1) IA estrecha, 2) IA general y 3) Súper Inteligencia Artificial (SIA). Clasificando por funcionamiento, podemos distinguir entre cuatro tipos: 1) Máquinas reactivas, 2) IA de memoria limitada, 3) IA de teoría de la mente, 4) IA autoconsciente. En los siguientes párrafos entraré a explicar lo que diferencia a cada una de ellas en detalle para situar al lector.

La IA estrecha[22] es un tipo de IA capaz de realizar con inteligencia una tarea muy concreta. Es el tipo de IA más común y actualmente disponible en el mundo de la IA. La IA estrecha no puede funcionar más allá de su angosto campo de conocimiento, ya que solo está entrenada para una tarea específica. Este tipo de IA puede fallar de manera impredecible si va más allá de sus límites.[23] Un ejemplo de IA estrecha son los sistemas de recomendación, utilizados por empresas como Amazon o Netflix.[24] El algoritmo de Netflix funciona como una combinación de aprendizaje automático y procesamiento del lenguaje natural, que recoge datos cómo las búsquedas que hace un usuario, desde que dispositivo se conecta, cuántos días y tiempo se conecta, etc. Utilizando estos datos, el algoritmo se va haciendo cada vez más eficiente. Prueba de la relevancia de tener un algoritmo bien optimizado es el hecho de que Netflix ofreció un premio por valor de 1

[21] Ramzai, J. (2020). *Clearly explained: 4 types of Machine Learning Algorithms*. Recuperado el 26 de marzo 2022. Towards Data Science: https://towardsdatascience.com/clearly-explained-4-types-of-machine-learning-algorithms-71304380c59a

[22] Ibídem.
[23] Ibídem.
[24] Ibídem.

millón de dólares en 2009 a quién consiguiese desarrollar un algoritmo para Netflix que fuera un 10% más efectivo.[25]

La IA general[26] es un tipo de inteligencia que puede realizar cualquier tarea intelectual con la misma eficacia que un ser humano. La idea detrás de la IA general es crear un sistema que pueda ser más inteligente y pensar como un humano por sí mismo. En la actualidad, no existe ningún sistema de este tipo, uno que pueda realizar cualquier tarea de forma tan perfecta como un ser humano.[27] El énfasis en este tipo de IA se encuentra en su capacidad para realizar una enorme variedad de tareas, incluso tareas a las que ni ha sido expuesta ni específicamente entrenada para hacer.[28] Es la siguiente frontera en el mundo de la IA y hay muchos investigadores alrededor del mundo tratando de construir este tipo de IA.[29]

La Súper Inteligencia Artificial[30] (SIA) es un nivel de Inteligencia de Sistemas en el que las máquinas superan la inteligencia humana y pueden realizar cualquier tarea que requiera esfuerzos cognitivos mejor que el ser humano. Es un resultado de la IA general (es IA general llevada al extremo, lo que hace que cambie su naturaleza fundamental, y por eso se crea otra categoría para distinguirla de la anterior). Algunas de las características clave de la SIA incluyen la capacidad de pensar, razonar, hacer juicios, planificar, aprender y comunicarse por sí misma.

A nivel teórico, los investigadores de Inteligencia Artificial no saben dar respuestas a preguntas cruciales para la SIA. Por ejemplo, no sabemos si podrán desarrollar sentimientos o códigos morales, principalmente, porque tampoco sabemos al completo lo que permite a los humanos crear códigos morales, ni entendemos de forma completa los procesos emocionales y su variedad individual dentro de la especie humana.[31] Es por ello que las Ciencias de la Computación (*Computer Science* en inglés)

[25] Van Buskirk, E. (2009). *BellKor's Pragmatic Chaos wins $1 million Netflix Prize by mere minutes.* Recuperado el 6 de diciembre de 2021. Wired: https://www.wired.com/2009/09/bellkors-pragmatic-chaos-wins-1-million-netflix-prize/
[26] Ramzai, J. (2020). *Clearly explained: 4 types of Machine Learning Algorithms*. Recuperado el 26 de marzo 2022. Towards Data Science: https://towardsdatascience.com/clearly-explained-4-types-of-machine-learning-algorithms-71304380c59a
[27] Ibídem.
[28] Ibídem.
[29] Ibídem.
[30] Ibídem.
[31] Ibídem.

están empezando a complementarse con áreas de conocimiento como la Neurociencia y la Psicología, tratando de descubrir patrones de funcionamiento en el cerebro que permitan hacer analogías en el desarrollo de sistemas de IA más potentes (cuyo máximo exponente sería la SIA). La SIA es todavía un concepto hipotético en el mundo de la IA, pero también uno de los que más atención recibe debido a sus potenciales implicaciones para la sociedad a gran escala.

Si pasamos a clasificar la IA según su funcionamiento, veremos que existen 4 tipos. De estos 4, el tipo más básico de IA son las llamadas "máquinas reactivas". Estos sistemas de IA no almacenan memorias o experiencias pasadas para utilizarlos en acciones futuras.[32] Sólo se centran en los escenarios actuales y reaccionan ante ellos, según la mejor acción posible. Un ejemplo de máquina reactiva sería el sistema Deep Blue de IBM diseñado para jugar al ajedrez (ganó al campeón mundial Garry Kasparov en 1997[33]). Deep Blue puede identificar las piezas de un tablero de ajedrez y saber cómo se mueve cada una. Puede hacer predicciones sobre los siguientes movimientos que debería hacer (y predicciones sobre los siguientes movimientos de su oponente), basándose en modelos probabilísticos. Así elige los movimientos óptimos entre las posibilidades que el tablero ofrece en una jugada concreta. Pero no tiene ningún concepto del pasado, ni ningún recuerdo de lo que ha sucedido antes. Aparte de una regla específica del ajedrez, raramente utilizada, que prohíbe repetir la misma jugada tres veces, Deep Blue ignora todo lo anterior al momento presente. Todo lo que hace es mirar las piezas en el tablero de ajedrez tal y como están en ese momento, y elegir entre los posibles movimientos siguientes.[34] Este tipo de inteligencia implica que el ordenador percibe el mundo directamente y actúa según lo que ve. No se basa en un concepto interno del mundo. Las máquinas reactivas se comportarán exactamente igual cada vez que se encuentren con la misma situación.[35] Esto puede ser francamente bueno para garantizar que un sistema de IA sea digno de confianza: es deseable que los coches autónomos

[32] Ketchell, M. (2016). *Understanding the 4 types of AI, from reactive robots to self-aware beings*. Recuperado el 8 de diciembre de 2021. The Conversation: https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616
[33] History.com Editors (2009). *Deep Blue defeats Garry Kasparov in chess match*. Recuperado el 8 de diciembre de 2021. HISTORY: https://www.history.com/this-day-in-history/deep-blue-defeats-garry-kasparov-in-chess-match

[34] Ketchell, M. (2016). *Understanding the 4 types of AI, from reactive robots to self-aware beings*. Recuperado el 8 de diciembre de 2021. The Conversation: https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616
[35] Ibídem.

actúen como un conductor fiable. Pero es inadecuado si nuestro objetivo es que las máquinas se relacionen realmente con el mundo y respondan a él.

Las máquinas de memoria limitada[36] pueden almacenar experiencias pasadas o algunos datos durante un corto periodo de tiempo. Estas máquinas pueden utilizar los datos almacenados sólo durante un periodo de tiempo limitado. Los coches autónomos se basan en este tipo de IA. Por ejemplo, observan la velocidad y la dirección de otros coches. Eso no puede hacerse en un solo momento, sino que requiere identificar objetos concretos y vigilarlos a lo largo del tiempo. Pero estas piezas de información son sólo transitorias. No se guardan como parte de la biblioteca de experiencias del coche, de la que puede aprender, del mismo modo que los conductores humanos acumulan experiencia durante años al volante.

La IA basada en la teoría de la mente[37] no sólo se forma representaciones sobre el mundo, sino también sobre otros agentes o entidades del mundo. En psicología, esto se denomina "teoría de la mente": la comprensión de que las personas, las criaturas y los objetos del mundo pueden tener pensamientos y emociones que afectan a su comportamiento. Esto permite a este tipo de IA predecir los comportamientos de los demás agentes basándose en lo que la IA *piensa* que están pensando.

Esta capacidad ha sido crucial para los seres humanos, permitiéndonos formar sociedades, tener interacciones sociales y desarrollar empatía cognitiva (entender lo que otra persona siente, aunque eso no cambie tus sentimientos, algo que tiene implicaciones para el comportamiento dependiendo del contexto y los agentes involucrados, entre otros.). Sin entender los motivos y las intenciones de los demás, y sin tener en cuenta lo que otra persona sabe sobre mí o sobre el entorno, trabajar juntos es difícil. Con el fin de tener sistemas de IA más integrados en la sociedad, muchos investigadores están concentrando sus esfuerzos en desarrollar este tipo de IA, aunque de momento no lo han conseguido.[38]

[36] Ibídem.
[37] Ibídem.
[38] Ibídem.

La IA autoconsciente[39] es el futuro de la Inteligencia Artificial. Estas máquinas serán súper inteligentes y tendrán sentimientos y autoconciencia. Serán más inteligentes que la mente humana (probablemente más inteligente que la combinación de toda la capacidad cognitiva de la Humanidad). El último paso de desarrollo en el mundo de la IA es construir sistemas que puedan formar representaciones sobre sí mismos.[40] Esto es, en cierto sentido, una extensión de la "teoría de la mente" que poseen el tipo de IA discutido en el párrafo anterior. Aunque probablemente estemos lejos de crear máquinas con conciencia de sí mismas, deberíamos centrar nuestros esfuerzos en comprender la memoria, el aprendizaje y la capacidad de basar las decisiones en experiencias pasadas.[41] Este es un paso importante para entender la inteligencia humana por sí misma. Y es crucial si queremos diseñar o hacer evolucionar máquinas que sean más que excepcionales que la inteligencia humana a la hora de clasificar lo que tienen delante.[42]

5.2 Tipos de aprendizaje automático

El aprendizaje automático ha pasado de ser un tema de ciencia ficción a ser la herramienta empresarial más fiable y diversa para optimizar los distintos aspectos de cada operación empresarial. Su impacto en el rendimiento de distintos negocios, en industrias que no tienen nada que ver entre sí (desde la Banca a las redes sociales), es tan significativa que la implementación de algoritmos de aprendizaje automático de primera categoría es necesaria para garantizar la supervivencia de muchas industrias.

La implementación del aprendizaje automático en las operaciones empresariales requiere la dedicación de una cantidad significativa de recursos y constituye una decisión estratégica importante para cualquier empresa. Ahora que entendemos por qué los algoritmos de aprendizaje automático son importantes, el siguiente paso es entender cómo las empresas pueden utilizarlos. El primer paso para ello es que la empresa entienda claramente cuál es el problema de negocio que le gustaría resolver utilizando algoritmos de aprendizaje automático; y entender claramente la cantidad de recursos y esfuerzos requeridos en los diferentes tipos de algoritmos de aprendizaje automático, para que

[39] Ibídem.
[40] Ibídem.
[41] Ibídem.
[42] Ibídem.

pueda elegir el algoritmo que mejor se adapte a su problema. Ahora pasaré a explicar los principales tipos de algoritmos de aprendizaje automático, el propósito de cada uno de ellos y sus potenciales beneficios a nivel empresarial.

Los principales tipos de algoritmos de aprendizaje automático son: 1) algoritmos de aprendizaje supervisado, 2) algoritmos de aprendizaje no supervisado, 3) algoritmos de aprendizaje semi-supervisado y 4) algoritmos de aprendizaje por refuerzo.[43]

Los algoritmos de aprendizaje supervisado son los más sencillos de los cuatro tipos de algoritmos de aprendizaje automático. Estos algoritmos requieren la supervisión directa del desarrollador del algoritmo.[44] En este tipo de algoritmos, el desarrollador etiqueta el corpus de datos muestrales (los datos obtenidos de una muestra de entre todos los datos existentes) y establece límites estrictos sobre los que operará el algoritmo.[45] Se trata de una versión del aprendizaje automático que "se alimenta con cuentagotas":

1. Se selecciona qué tipo de información de salida (muestras) se va a "alimentar" al algoritmo.
2. Se determina manualmente el tipo de resultados deseados (por ejemplo, "sí/no" o "verdadero/falso").

El objetivo principal del aprendizaje supervisado es ampliar el alcance de los datos y hacer predicciones de datos no disponibles, futuros o no vistos, sirviéndose de datos muestrales etiquetados. El aprendizaje automático supervisado se basa principalmente en dos procesos: 1) clasificación y 2) regresión.[46] La clasificación es el proceso de aprendizaje a partir de muestras de datos pasadas y el entrenamiento manual del modelo para predecir resultados esencialmente binarios (sí/no, verdadero/falso, 0/1). Por ejemplo: si un empleado va a ser despedido en el próximo año o no, si alguien tiene una enfermedad o no, etc. El algoritmo de clasificación reconoce ciertos tipos de objetos y los clasifica en consecuencia, para predecir uno de los dos resultados posibles. La regresión es el proceso

[43] Ramzai, J. (2020). *Clearly explained: 4 types of Machine Learning Algorithms*. Recuperado el 26 de marzo 2022. Towards Data Science: https://towardsdatascience.com/clearly-explained-4-types-of-machine-learning-algorithms-71304380c59a
[44] Ibídem.
[45] Ibídem.
[46] Ibídem.

de identificar patrones y calcular las predicciones de resultados de naturaleza continua (como el peso de una persona, la previsión de ventas de este año en una empresa, etc.) en vez de binarios.

Los campos de uso más comunes para algoritmos de aprendizaje supervisado son la predicción de precios y la previsión de tendencias en las ventas, el comercio minorista y el *trading* de acciones.[47] Estos algoritmos utilizan datos entrantes para evaluar las probabilidades de diferentes escenarios y calcular así los resultados (predicciones).

Los algoritmos de aprendizaje no supervisado no implican control directo por parte del desarrollador.[48] El requisito fundamental para utilizar algoritmos de aprendizaje automático supervisado es que debemos conocer los resultados de los datos pasados de antemano, para así poder predecir los resultados en los datos no vistos, a diferencia de lo que ocurre con los algoritmos de aprendizaje automático no supervisado, donde los resultados deseados son desconocidos y aún están por definir.[49]

La distinta forma de funcionar de los distintos tipos de algoritmos se debe a que sirven para resolver problemas de naturaleza distinta.[50] Hay ocasiones en las que no se quiere predecir exactamente un resultado (tarea para la cual utilizaríamos un algoritmo de aprendizaje supervisado), sino que se quiere realizar una segmentación o agrupación de resultados. Por ejemplo, un banco quiere tener una segmentación de sus clientes para entender su comportamiento mejor, y así modificar la forma en la que ofrece servicios bancarios. Este problema de negocio requiere el uso de algoritmos de aprendizaje automático no supervisado, ya que aquí no se predicen resultados específicos, sino que se está utilizando un algoritmo para mapear o descubrir la información que esa agrupación de datos nos aporta sobre los clientes.

En definitiva, se usa un algoritmo de aprendizaje automático no supervisado cuando se desea resolver un problema, pero no se sabe exactamente su raíz o la mejor forma de optimizarlo. Se aplica el algoritmo a una gran cantidad de datos, se realizan

[47] Ibídem.
[48] Ibídem.
[49] Ibídem.
[50] Ibídem.

agrupaciones y así se descubren los patrones existentes en los datos, que servirán para delimitar una estrategia acorde con el problema objeto de análisis. Otra gran diferencia entre ambos tipos de algoritmos es que el aprendizaje supervisado utiliza exclusivamente datos etiquetados, mientras que el aprendizaje no supervisado se alimenta de datos no etiquetados.

El algoritmo de aprendizaje automático no supervisado se utiliza para explorar la estructura de la información, extraer ideas valiosas; detectar patrones, con el objetivo de aumentar la eficiencia (independientemente del área en que se esté aplicando el algoritmo).

Los algoritmos de aprendizaje no supervisado aplican las siguientes técnicas[51] para describir los datos:

- Agrupación: consiste en una exploración de los datos que se utiliza para segmentarlos en grupos significativos (llamados *clusters*) basados en sus patrones internos, sin ningún conocimiento previo de las categorías según las cuales se crean esos *clusters*. Las categorías se definen por la similitud existente entre un grupo de datos individuales y al mismo tiempo por patrones en su distinción con el resto de datos individuales no pertenecientes al *cluster* (que también pueden utilizarse para detectar anomalías). Es decir, la agrupación tiene lugar vía positiva (similitud) y al mismo tiempo vía negativa (diferencia con el resto de datos).
- Reducción de la dimensionalidad: En la mayoría de casos objeto de análisis por parte de algoritmos de aprendizaje automático no supervisado, hay demasiado "ruido" en los datos entrantes. El problema de ruido vs señal es muy importante de resolver para la ciencia de datos (aunque también es aplicable a nivel humano). La señal es la información significativa que se intenta detectar. El ruido es la variación o fluctuación aleatoria y no deseada que interfiere con la señal. Los algoritmos de aprendizaje automático utilizan la reducción de la dimensionalidad para eliminar este ruido y destilar la información relevante (encontrar la señal).

[51] Ibídem.

El uso de algoritmos de aprendizaje automático no supervisado se extiende a una multitud de industrias. Los principales sectores[52] donde se utilizan este tipo de algoritmos son: 1) el marketing digital, donde se utilizan para identificar grupos de público objetivo en función de ciertas categorías (pueden ser datos de comportamiento, elementos de datos personales, configuración de software específica...) y 2) la tecnología publicitaria, donde estos algoritmos pueden utilizarse para desarrollar una orientación más eficiente del contenido de los anuncios y para identificar patrones en el rendimiento de las campañas. También se utilizan cuando es necesario explorar información del cliente para así adaptar los servicios pertinentes.

Los algoritmos de aprendizaje automático semi-supervisado son un punto intermedio entre los algoritmos supervisados y los no supervisados. En esencia, el algoritmo semi-supervisado combina algunos aspectos de ambos para convertirse en algo particular, que ninguno de los otros dos tipos de algoritmos puede hacer por si solo.

El aprendizaje semi-supervisado utiliza el proceso de clasificación para identificar datos que pueden considerarse activos para la empresa, mientras el proceso de *clustering* hace agrupaciones para poder inferir la máxima información relevante de esos datos previamente clasificados.[53] Los sectores jurídico y sanitario, entre otros, gestionan la clasificación de contenidos web, la clasificación de imágenes y el análisis del habla con la ayuda del aprendizaje semi-supervisado.[54] En el caso de la clasificación de contenidos web, el aprendizaje semi-supervisado se aplica a los motores de rastreo y a los sistemas de agregación de contenido. En ambos casos, utiliza una amplia gama de etiquetas para analizar los contenidos y ordenarlos en configuraciones específicas. Sin embargo, este procedimiento suele requerir la intervención humana para su posterior clasificación.

El aprendizaje por refuerzo se conoce también por el nombre de aprendizaje reforzado, y es el tipo de aprendizaje automático comunmente conocido como Inteligencia Artificial. Esencialmente, el aprendizaje por refuerzo consiste en desarrollar un sistema autosuficiente que, a lo largo de secuencias continuas de prueba y error, se mejora a sí mismo basándose en la combinación de datos etiquetados y en las

[52] Ibídem.
[53] Ibídem.
[54] Ibídem.

interacciones con los datos entrantes.[55] El aprendizaje automático por refuerzo se adapta a casos donde la información disponible es limitada o inconsistente. En este caso, un algoritmo puede formar sus procedimientos operativos basándose en las interacciones iterativas con los datos y los procesos relevantes.

Los coches autónomos se basan en algoritmos de aprendizaje por refuerzo[56], haciéndose cada vez mejores cuantas más horas de conducción tienen (ya que les ha dado tiempo a analizar millones de imágenes y compararlas con las decisiones tomadas durante la conducción). Por ejemplo, si un coche autónomo detecta que la carretera gira a la izquierda, puede activar el escenario "girar a la izquierda" y así sucesivamente. Por otro lado, las operaciones de Marketing y Ad-Tech también utilizan el aprendizaje por refuerzo. Este tipo de algoritmo de aprendizaje automático puede hacer que las operaciones de *retargeting* sean mucho más flexibles y eficientes a la hora de conseguir la conversión de los clientes, adaptándose estrechamente al comportamiento del usuario y al contexto que le rodea.[57]

Además, el aprendizaje por refuerzo se utiliza para ampliar y ajustar el procesamiento del lenguaje natural (NLP) y la generación de diálogos en los *chatbots*[58], con el fin de: 1) imitar el estilo de un mensaje entrante, 2) desarrollar respuestas más atractivas e informativas, 3) encontrar respuestas relevantes según la reacción del usuario.[59]

5.3 Situación actual de las investigaciones sobre ética de los algoritmos

Determinar el impacto ético potencial y real de un algoritmo es difícil por muchas razones. Identificar la influencia humana en el diseño y la configuración de los algoritmos suele requerir la investigación de procesos de desarrollo a largo plazo y con múltiples usuarios. Incluso con recursos suficientes, los problemas y los valores subyacentes a menudo no serán evidentes hasta que surja un caso de uso problemático en el mundo real.

[55] Ibídem.
[56] Ibídem.
[57] Ibídem.
[58] Ibídem.
[59] Ibídem.

Los algoritmos de aprendizaje, a menudo citados como el "futuro" de los algoritmos y la analítica[60], introducen incertidumbre sobre cómo y por qué se toman las decisiones debido a su capacidad para modificar los parámetros operativos y las reglas de toma de decisiones por sí mismos, en el proceso de estar siendo aplicados[61]. Determinar si una decisión problemática concreta es simplemente un error puntual o prueba de un error sistémico puede ser imposible (o al menos muy difícil) con algoritmos de aprendizaje poco interpretables y predecibles. Este tipo de problemas se incrementará a medida que la complejidad de los algoritmos aumente, ya que llegará un momento en que los algoritmos interactúen con los resultados de los demás algoritmos para tomar decisiones[62]. Por ello, es importante resaltar cual es el estado general de este nicho dentro del ámbito académico de la IA.

Desde hace años se viene investigando el aspecto ético del desarrollo y aplicación de algoritmos, pero dada la complejidad y el vertiginoso desarrollo de nuevos tipos de algoritmo ha impedido una visión unificada de ellos desde una perspectiva ética. A pesar de esto, la mayoría de la literatura está de acuerdo en dos puntos fundamentales.[63] En primer lugar, están de acuerdo con que los algoritmos son, intrínsecamente, herramientas imbuidas de valor[64]/juicios de valor, ya que la forma en que los desarrolladores los crean especificando unos parámetros operacionales, y la forma en que los usuarios los configuran, da a entender que, por necesidad, están buscando priorizar unos valores o intereses sobre otros. En segundo lugar, los investigadores están de acuerdo en que el hecho de que el algoritmo opere dentro de los parámetros que se le han dado no garantiza resultados éticamente aceptables[65].

A través de un análisis de mi revisión de la literatura, se hace aparente que los algoritmos presentan desafíos éticos por varias razones: 1) su escala, siendo herramientas que impactan la vida de hasta miles de millones de personas, 2) la opacidad/incertidumbre

[60] Tutt, A. (2017). An FDA for algorithms. *Admin. L. Rev.*, *69*, 83.
[61] Burrell J (2016) How the machine 'thinks:' Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.
[62] Tutt, A. (2017). An FDA for algorithms. *Admin. L. Rev.*, *69*, 83.
[63] Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of business ethics*, *160*(4), 835-850.
[64] Brey P. y Soraker J.H. (2009). Philosophy of Computing and Information Technology. *Elsevier*.
[65] Sweeney L. (2013). Discrimination in online ad delivery. *Queue* 11(3): 10:10–10:29.

sobre el funcionamiento del algoritmo (problema especialmente acuciante con los algoritmos basados en aprendizaje automático), y 3) su impacto (muchos algoritmos deciden sobre cosas importantes en la vida de una persona, como la idoneidad para recibir un crédito o no) me parecen las razones más importantes de destacar. Sin embargo, estos no son más consideraciones o efectos de segundo orden, derivados de una conceptualización ética inadecuada.

Con el fin de realizar un análisis lo más "denso" posible, he decidido utilizar el que, personalmente, me parece el mejor marco conceptual desde el que analizar las implicaciones éticas de los algoritmos, un mapa[66] desarrollado por Mittlestadt et al. El mapa propuesto puede utilizarse para organizar el discurso académico actual que describe las preocupaciones éticas sobre los algoritmos fundamentadas sobre bases éticas y epistémicas. Siguiendo el mapa propuesto por Mittlestadt, podemos analizar la ética de los algoritmos de acuerdo con 6 tipos de cuestiones éticas a resolver: tres cuestiones epistémicas, dos cuestiones normativas, y una cuestión ética más global. Son las siguientes: 1) evidencia no concluyente, 2) evidencia inescrutable, 3) evidencia equivocada, 4) resultados injustos, 5) efectos transformadores y 6) trazabilidad. Este mapa centra una de las preguntas más importantes de la investigación: ¿qué tipo de cuestiones éticas plantean los algoritmos? Sabiendo la respuesta a esa pregunta, el camino para delimitar la responsabilidad de cada *stakeholder* será más certero, algo que se verá ejemplificado en la aplicación práctica del mapa al análisis de la responsabilidad en Facebook.

5.4 Puntos clave y tendencias en el ámbito de investigación de la ética de los algoritmos

Los puntos clave en el ámbito del uso ético de algoritmos son: 1) la responsabilidad distribuida, 2) el mal funcionamiento de los algoritmos (podemos interpretar los algoritmos con consecuencias imprevistas negativas como algoritmos que funcionan mal), 3) la forma en la que se cumple con el requisito de transparencia, 4) la

[66] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*.

mejora de mecanismos regulatorios.[67] A lo largo de este epígrafe hablaremos más a fondo de cada uno de ellos.

Mi intención en el epígrafe anterior ha sido describir el estado del discurso académico en torno a la ética de los algoritmos y proponer una herramienta de organización para el análisis ético que llevaré a cabo en este trabajo. El mapa puede resultar una herramienta útil, que mejore la descripción de los problemas éticos de los algoritmos, y permita analizarlos de forma más adecuada. Como indica el mapa, las preocupaciones éticas de los algoritmos son multidimensionales y, por tanto, requieren soluciones multidimensionales.[68]

Aunque el mapa proporciona la estructura conceptual básica que necesitamos, aún debe poblarse a medida que proliferan el despliegue y los estudios críticos de algoritmos. Teniendo esto en cuenta, en esta sección planteamos una serie de temas que aún no han recibido una atención sustancial en la literatura revisada relacionados con 1) los efectos transformadores y 2) la trazabilidad de los algoritmos.[69] Estos temas pueden considerarse futuras tendencias de investigación para la ética de los algoritmos a medida que el campo se expande y madura.[70]

En cuanto a los efectos transformadores, los algoritmos cambian la forma de construir, gestionar y proteger la identidad del usuario por parte de los mecanismos de privacidad y protección de datos.[71] La privacidad informativa y la posibilidad de ser identificable suelen estar estrechamente vinculadas; un individuo tiene privacidad en la medida en que tiene control sobre sus datos e información. En la medida en que los algoritmos transforman la privacidad, haciendo que la posibilidad de ser identificable sea menos importante, requerimos de una teoría de la privacidad que responda a la menor importancia de la posibilidad de ser identificable y de la individualidad.[72]

[67] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.
[68] Ibídem.
[69] Ibídem.
[70] Ibídem.
[71] Ibídem.
[72] Ibídem.

Van Wel y Royakkers instan a una nueva conceptualización de los datos personales[73], en la que se concedan protecciones de privacidad equivalentes a las "características de grupo", cuando estas características se utilicen en lugar de las "características individuales" a la hora de generar conocimiento sobre un individuo o de emprender acciones respecto a él. Para que esto funcione, es necesario seguir trabajando para describir cómo la privacidad opera a nivel de "características de grupo" en ausencia de la posibilidad del usuario de ser identificable (a nivel individual).

También es necesario desarrollar mecanismos utilizables en el mundo real que hagan eficaz ese cumplimiento de protección de la privacidad en el uso del Big Data. No debemos dar por sentado que los algoritmos discriminen según características observables o comprensibles para los humanos. Por ello, es necesario desarrollar otros mecanismos de detección de daños, dada la capacidad de los algoritmos para perjudicar a los usuarios de forma indirecta y no evidente, muchas veces superando las definiciones legales de discriminación[74].

En lo que respecta a la trazabilidad, siguen existiendo dos retos fundamentales para la responsabilidad de los algoritmos. En primer lugar, a pesar de la abundante literatura que aborda la responsabilidad moral y la agencia propia de los algoritmos, no se ha prestado suficiente atención a la responsabilidad distribuida, es decir, la responsabilidad compartida por una combinación de actores humanos y algorítmicos simultáneamente[75]. La literatura revisada aborda la posible agencia moral de los algoritmos, pero no describe los métodos y principios para repartir la culpa o la responsabilidad en una red mixta de actores humanos y algorítmicos[76].

En segundo lugar, ya se ha depositado una gran confianza en los algoritmos, en algunos casos desembocando en una evasión de responsabilidad por parte de los actores humanos, o una tendencia a "esconderse detrás del ordenador" y asumir que los procesos

[73] Van Wel L and Royakkers L. (2004). Ethical issues in web data mining. *Ethics and Information Technology* 6(2): 129–140.
[74] Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*.
[75] Simon J (2015) Distributed epistemic responsibility in a hyperconnected era. In: Floridi L (ed.) *The Onlife Manifesto*. Springer International Publishing, pp. 145–159.
[76] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*. pp.12.

automatizados son correctos por defecto[77]. Delegar la toma de decisiones a los algoritmos puede alejar la responsabilidad de los responsables humanos. Los algoritmos en los que participan actores de múltiples disciplinas pueden, por ejemplo, llevar a que cada parte asuma que las otras partes asumirán la responsabilidad ética por las acciones del algoritmo[78], dando lugar a una evasión de responsabilidad en la práctica.

El aprendizaje automático añade una capa adicional de complejidad entre los desarrolladores y las acciones tomadas por el algoritmo, lo que, de forma justificada, puede debilitar la responsabilidad de los desarrolladores. Se necesita más investigación para comprender la prevalencia de estos efectos en los sistemas de toma de decisiones basados en algoritmos, y para discernir cómo distribuir la responsabilidad en esos casos de justificación derivados de la complejidad de los algoritmos basados en aprendizaje automático, cuando sus decisiones resultan en perjuicio de alguien.

Un problema relacionado es el del mal funcionamiento. La necesidad de repartir la responsabilidad se hace sentir de forma aguda cuando los algoritmos funcionan mal. Resulta útil distinguir entre: 1) errores de diseño (tipos) y 2) errores de funcionamiento (tokens), y entre la falta de funcionamiento esperado (disfunción) y la presencia de efectos secundarios no deseados (mal funcionamiento)[79]. La disfunción se distingue de los meros efectos secundarios negativos por la posibilidad de evitarlos, o por la medida en que tipos comparables de algoritmos cumplen la función prevista sin los efectos negativos.[80] Estas distinciones aclaran los aspectos éticos de los algoritmos que están estrictamente relacionados con su funcionamiento, ya sea en abstracto, o como parte de un sistema de toma de decisiones más amplio, y revela la interacción polifacética entre el comportamiento previsto y el real.[81]

Ambos tipos de "mal funcionamiento" implican distintas responsabilidades para los desarrolladores de algoritmos, las empresas y los usuarios. Hay que seguir trabajando

---

[77] Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.

[78] Davis M, Kumiega A and Van Vliet B. (2013). Ethics, finance, and automation: A preliminary survey of problems in high frequency trading. *Science and Engineering Ethics* 19(3): 851–874.

[79] Floridi L, Fresco N and Primiero G (2014) On malfunctioning software. *Synthese* 192(4): 1199–1220.

[80] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

[81] Ibídem.

para describir el reparto equitativo de la responsabilidad por disfunción y mal funcionamiento en grandes equipos de desarrollo y en contextos complejos de uso.[82] También es necesario seguir trabajando para especificar los requisitos de resistencia al mal funcionamiento como un ideal ético en el diseño de algoritmos.[83] El aprendizaje automático, en particular, plantea retos únicos, porque lograr un comportamiento previsto o "correcto" no implica la ausencia de errores[84] o de bucles de retroalimentación perjudiciales. Los algoritmos pueden, por ejemplo, hacerse interrumpibles de forma segura, en la medida en que se pueden desalentar las acciones perjudiciales derivadas del algoritmo, sin que el algoritmo se vea animado a engañar a los usuarios en el futuro para evitar más interrupciones[85]. Quizá este debería ser un diseño de base en todos los algoritmos con impacto relevante a nivel empresarial.

Por último, si bien se reconoce ampliamente que es necesario algo de transparencia respecto al algoritmo, como requisito para su trazabilidad, la forma de transparencia sigue siendo una cuestión abierta, en particular para el aprendizaje automático. El mero hecho de hacer transparente el código de un algoritmo no es suficiente para garantizar su comportamiento ético. Los requisitos normativos o metodológicos que se han impuesto de que los algoritmos sean *explicables* o *interpretables* demuestran el reto al que se enfrentan ahora los procesadores de datos[86].

Un posible camino hacia la explicabilidad es la auditoría algorítmica llevada a cabo por los procesadores de datos[87], reguladores externos[88], o investigadores

[82] Ibídem.
[83] Ibídem.
[84] Burrell, J. (2016). How the machine 'thinks:' Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.
[85] Orseau, L. y Armstrong, S. (2016). *Safely interruptible agents*. [Archivo PDF]. URL: http://intelligence.org/files/Interruptibility.
[86] Tutt, A. (2017). An FDA for algorithms. *Admin. L. Rev., 69*, 83.
[87] Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.
[88] Pasquale, F. (2015). *The black box society*. Harvard University Press.

empíricos[89], utilizando estudios de auditoría a posteriori[90] o mecanismos de información diseñados en el propio algoritmo[91].

Para todos los tipos de algoritmos, la auditoría es una condición previa necesaria para verificar su correcto funcionamiento. Para los algoritmos de aprendizaje automático con impacto humano previsible, la auditoría puede crear un registro procesal a posteriori de la toma de decisiones algorítmicas, para desentrañar decisiones problemáticas o inexactas, o para detectar la discriminación o daños similares[92]. Es necesario seguir trabajando para diseñar mecanismos de auditoría ampliamente aplicables a distintos tipos de algoritmo[93], basados en los mecanismos actuales de transparencia e interpretabilidad de algoritmos de aprendizaje automático[94].

Un último pero importante aspecto que requiere más trabajo es la configuración de normativas y mecanismos de regulación realistas**.** El Reglamento General de Protección de Datos de la UE (RGPD) es indicativo de los retos a los que se enfrentará la regulación de los algoritmos en todo el mundo.

El RGPD estipula una serie de responsabilidades de los responsables del tratamiento de los datos (en su mayor parte empresas), además de derechos de los usuarios cuyos datos son recabados por empresas, y utilizados por estas a través de algoritmos, buscando mejorar la toma de decisiones empresariales. Cuando los responsables del tratamiento de datos realicen la elaboración de perfiles, deberán evaluar las posibles consecuencias de sus actividades de tratamiento de datos a través de una evaluación de impacto sobre la protección de datos (art. 35.3.a). Además de evaluar los riesgos para la privacidad, los responsables del tratamiento de datos también tienen que comunicar estos

---

89 Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, communication & society*, *20*(1), 14-29.

90 Adler P, Falk C, Friedler SA, et al. (2016) Auditing black-box models by obscuring features. arXiv:1602.07043 [cs, stat]. URL: http://arxiv.org/abs/1602.07043

91 Vellido, A., Martín-Guerrero, J. D., & Lisboa, P. J. (2012). Making machine learning models interpretable. In *ESANN* (Vol. 12, pp. 163-172).

92 Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*. pp.13.

93 Adler P, Falk C, Friedler SA, et al. (2016) Auditing black-box models by obscuring features. arXiv:1602.07043 [cs, stat]. URL: http://arxiv.org/abs/1602.07043

94 Kim, B., Patel, K., Rostamizadeh, A., & Shah, J. (2015). Scalable and interpretable data representation for high-dimensional, complex data. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1).

riesgos a las personas afectadas. De acuerdo con los arts. 13.2f) y 14.2g), los responsables del tratamiento están obligados a informar a los interesados sobre los métodos de elaboración de perfiles, su significancia y sus consecuencias previstas. El art. 12.1 exige que se utilice un lenguaje claro y sencillo para informar sobre estos riesgos. Además, el Considerando 71 establece la obligación de los responsables del tratamiento de explicar la lógica de cómo se ha llegado a la decisión. Por último, el art. 22.3 establece la obligación del responsable del tratamiento de "aplicar las medidas adecuadas para salvaguardar los derechos y libertades del interesado y sus intereses legítimos" cuando se aplique la toma de decisiones automatizada. Esta obligación es bastante vaga y opaca.

En cuanto a los derechos de los usuarios, el RGPD adopta un enfoque de autodeterminación. Los interesados tienen derecho a oponerse a los métodos de elaboración de perfiles (art. 21) y el derecho a no ser sometidos a un proceso de toma de decisiones individuales exclusivamente automatizado (art. 22). El Considerando 71 explica que la toma de decisiones individuales exclusivamente automatizada debe entenderse como un método "que produce efectos jurídicos sobre él o ella o que le afecta de manera significativa, como la denegación automática de una solicitud de crédito en línea o las prácticas de contratación electrónica sin ninguna intervención humana" e incluye la elaboración de perfiles que permite "predecir aspectos relativos al rendimiento laboral, situación económica, salud, preferencias o intereses personales, fiabilidad o comportamiento, ubicación o movimientos". En estos y otros casos similares, el interesado tiene derecho a oponerse a la utilización de dichos métodos o, al menos, a obtener la intervención de una persona para expresar su opinión y para "impugnar la decisión" (art. 22.3).

A primera vista, estas disposiciones atribuyen el control a los interesados y les permiten decidir cómo se utilizan sus datos. A pesar de que el RGPD ha mejorado en gran medida la protección de datos, una serie de exenciones limitan los derechos de los interesados. El RGPD puede ser un mecanismo ineficaz o poderoso[95] de proteger a los interesados, dependiendo de su eventual interpretación jurídica: la redacción del reglamento permite que cualquiera de las dos cosas sea cierta. Las autoridades legislativas y judiciales, con sus futuras sentencias, determinarán la eficacia del RGPD. Sin embargo,

[95] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

se requiere trabajo adicional para proporcionar directrices, normativas y mecanismos prácticos que permitan poner en práctica los nuevos derechos y responsabilidades.

6. ANÁLISIS ÉTICO DE LA RESPONSABILIDAD EN LA EMPRESA DERIVADA DEL USO DE ALGORITMOS

6.1 Evidencia no concluyente

Cuando los algoritmos sacan conclusiones de los datos que procesan utilizando estadísticas inferenciales y/o técnicas de aprendizaje automático, producen un conocimiento probable (en el sentido de relacionado con la probabilidad y la estadística) pero, inevitablemente, conocimiento incierto. La teoría del aprendizaje estadístico[98] y la teoría del aprendizaje computacional[99] se ocupan de la caracterización y cuantificación de esta incertidumbre característica del conocimiento obtenido a través de algoritmos. Además de esto, los métodos estadísticos pueden ayudar a identificar correlaciones significativas, pero rara vez se consideran suficientes para que exista una relación causal[100] entre esas variables correlacionadas, y, por tanto, esas correlaciones significativas pueden ser insuficientes para motivar la acción basándose en la "certeza" de existencia de dicha relación causal. De ahí el énfasis que pone el mundo del Big Data sobre los *actionable insights* (en español "conocimientos procesables"). Ese énfasis en la practicidad, por encima de la certeza teórica del conocimiento, puede considerarse como un reconocimiento explícito de las limitaciones epistémicas, derivadas de los propios métodos estadísticos utilizados en el desarrollo de algoritmos.

Los algoritmos suelen utilizarse en contextos en los que técnicas más fiables no están disponibles o son demasiado caras (respecto al binomio coste/beneficio de utilizar el algoritmo), por lo que rara vez se pretende que sean infalibles. Desde un punto de vista empresarial, el algoritmo es una herramienta que busca mejorar la toma de decisiones, no crear un sistema de decisiones perfecto (aunque podemos conjeturar que ese es el objetivo final lógico de este desarrollo). Reconocer esta limitación es importante, pero debe

[98] James G, Witten D, Hastie T, et al. (2013) *An Introduction to Statistical Learning*. Vol. 6, New York: Springer.
[99] Valiant LG (1984) A theory of the learnable. Communications of the Journal of the ACM 27: 1134–1142.
[100] Illari PM and Russo F (2014) Causality: Philosophical Theory Meets Scientific Practice. Oxford: Oxford University Press.

complementarse con una evaluación de cómo el riesgo de equivocarse afecta a las propias responsabilidades epistémicas[101]: por ejemplo, debilitando la justificación que uno tiene para haber llegado a una conclusión (opinión) sobre un tema determinado, más allá de lo que se consideraría aceptable para justificar la acción en determinado contexto.

Gran parte de la toma de decisiones algorítmica y de la minería de datos se basa en el conocimiento inductivo y las correlaciones significativas identificadas dentro de un conjunto de datos. La causalidad no puede establecerse hasta haber actuado sobre la evidencia producida por el algoritmo. La búsqueda de relaciones causales entre dos variables es difícil, ya que las correlaciones entre variables, establecidas en grandes conjuntos de datos propiedad de una empresa, en la mayoría de los casos no son reproducibles ni falsables[102], ergo, su certeza no puede ser discernida. A pesar de esto, las correlaciones basadas en un conjunto de datos lo suficientemente grande se consideran cada vez más creíbles como base de conocimiento para motivar una determinada acción, sin necesidad de haber establecido primero la certeza de la relación causal[103]. En este sentido, la minería de datos y la elaboración de perfiles de datos a menudo solo necesita establecer una base de evidencias lo suficientemente fiable como para impulsar la acción, lo que aquí se denomina como "conocimientos procesables", en lugar de una base de evidencias completa y cierta.

Actuar sobre las correlaciones puede ser doblemente problemático. Sin embargo, hay que distinguir entre la justificación ética de actuar sobre la base de una mera correlación, y una ética más amplia del razonamiento inductivo que se superpone con las críticas existentes a los métodos estadísticos y cuantitativos en la investigación. La primera se refiere a los umbrales de evidencia requeridos para justificar acciones con impacto ético. La segunda se refiere a la falta de reproducibilidad del conocimiento en el campo del Big Data, que, en la práctica, distingue a este campo de la Ciencia[104], y se entiende mejor como una cuestión de epistemología.

---

[101] Miller B and Record I (2013) Justified belief in a digital age: On the epistemic implications of secret Internet technologies. *Episteme* 10(2): 117–134.

[102] Lazer D, Kennedy R, King G, et al. (2014) The parable of Google flu: Traps in big data analysis. *Science* 343(6176): 1203–1205.

[103] Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.

[104] Ioannidis JPA (2005) Why most published research findings are false. *PLoS Medicine* 2(8): e124.

Pueden descubrirse correlaciones espurias en lugar de verdadero conocimiento causal. En el análisis predictivo, las correlaciones son doblemente inciertas[105]. Incluso si se descubren de verdad correlaciones fuertes o conocimientos causales, este conocimiento puede afectar únicamente a las poblaciones (que es el objeto de estudio de los métodos estadísticos), mientras que las acciones se dirigen a los individuos[106], ergo, los conocimientos derivados estadísticamente no son necesariamente extrapolables a los individuos. "Las categorías algorítmicas... señalan la certeza, desalientan la exploración y crean coherencia entre objetos dispares"[107], todo lo cual contribuye a que los individuos sean vistos (posiblemente de forma inexacta) a través de modelos simplificados[108].

Por último, aunque tanto las acciones como el conocimiento se encuentran a nivel de la población, las acciones de una empresa pueden extenderse al nivel individual. Vemos la diatriba ética que se presenta en los casos empresariales de, por ejemplo, las aseguradoras, cuando estas fijan una prima de seguro (derivada de un algoritmo entrenado mediante métodos estadísticos) para una determinada subpoblación y, por tanto, tiene que ser pagada por cada miembro, independientemente de que no necesariamente el "conocimiento" que el algoritmo ha inferido de la subpoblación a la que pertenece el individuo sea atribuible a este. Las medidas adoptadas sobre la base de correlaciones inductivas tienen un impacto real en los intereses humanos, independientemente de su validez, y esta es una cuestión ética que es importante resolver.

6.2 Evidencia inescrutable y opacidad

La viabilidad de escrutar la evidencia, evaluada en términos de la transparencia u opacidad de los algoritmos, ha resultado ser una preocupación en la literatura revisada. La transparencia normalmente es deseable porque los algoritmos que son poco predecibles o explicables son difíciles de controlar, supervisar y corregir[109]. Muchos

[105] Ananny M (2016) Toward an ethics of algorithms convening, observation, probability, and timeliness. *Science, Technology & Human Values* 41(1): 93–117.
[106] Illari PM and Russo F (2014) Causality: Philosophical Theory Meets Scientific Practice. Oxford: Oxford University Press.
[107] Ananny M (2016) Toward an ethics of algorithms convening, observation, probability, and timeliness. *Science, Technology & Human Values* 41(1): 93–117. La frase original es: "algorithmic categories... signal certainty, discourage alternative explorations, and create coherence among disparate objects"
[108] Barocas, S. (2014). Data mining and the discourse on discrimination. In *Data ethics workshop, conference on knowledge discovery and data mining* (pp. 1-4).
[109] Tutt, A. (2017). An FDA for algorithms. *Admin. L. Rev.*, *69*, 83.

críticos[110], han observado que la transparencia es a menudo tratada (ingenuamente) como un elixir para problemas éticos que surgen de las nuevas tecnologías. La transparencia se define generalmente respecto a "la disponibilidad de la información, las condiciones de accesibilidad y el modo en que la información... puede apoyar pragmática o epistémicamente el proceso de toma de decisiones del usuario"[111]. El debate sobre este tema no es nuevo. La literatura sobre ética de la información y la informática, por ejemplo, empezó a centrarse en él a principios del siglo XXI, cuando empezaron a surgir las cuestiones relativas a la manipulación (en el sentido de ordenar los resultados de forma distinta según las distintas subpoblaciones a las que pertenece un usuario) de información por parte de los motores de búsqueda.

Los principales componentes de la transparencia son 1) la accesibilidad y 2) la comprensibilidad de la información. La información sobre la funcionalidad de los algoritmos a menudo es poco accesible, y se diseña así de forma intencionada. Los algoritmos patentados se mantienen en secreto en aras de la ventaja competitiva[112], la seguridad nacional[113], o la privacidad. Por tanto, la transparencia puede ir en contra de otros ideales éticos, en particular la libre competencia entre empresas y la autonomía de las organizaciones.

Granka señala una lucha de poder[114] entre los intereses de los usuarios, interesados en la transparencia, y la viabilidad comercial de los procesadores de datos, es decir, de las empresas que capturan esos datos y diseñan algoritmos para optimizar sus operaciones empresariales. La divulgación de la estructura de estos algoritmos facilitaría ver las manipulaciones malintencionadas de los resultados de búsquedas, sin que ello suponga ninguna ventaja para el usuario medio no experto en tecnología[115]. La viabilidad comercial de los procesadores de datos en muchos sectores (p. ej. entidades de crédito,

[110] Raymond A (2014) The dilemma of private justice systems: Big Data sources, the cloud and predictive analytics. *Northwestern Journal of International Law & Business, Forthcoming*. URL: http://papers.ssrn.com/sol3/ papers.cfm?abstract_id=2469291 Recuperado el 27 de diciembre 2021.
[111] Turilli M and Floridi L (2009) The ethics of information transparency. *Ethics and Information Technology* 11(2): 105–112.
[112] Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, communication & society*, *20*(1), 14-29.
[113] Leese M (2014) The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue* 45(5): 494–511.
[114] Granka LA (2010) The politics of search: A decade retrospective. *The Information Society* 26(5): 364–374.
[115] Granka LA (2010) The politics of search: A decade retrospective. *The Information Society* 26(5): 364–374.

trading de alta frecuencia, redes sociales) puede verse amenazada por la transparencia. Sin embargo, los usuarios de los datos conservan interés en comprender cómo se crea la información sobre ellos y cómo esta influye en las decisiones empresariales, ya que les afectan directamente. Esta lucha está marcada por la asimetría de la información y un "desequilibrio de conocimientos y poder de decisión" que favorece a los procesadores de datos[116].

Además de ser accesible, la información debe ser comprensible para ser considerada transparente[117]. Los esfuerzos por hacer que los algoritmos sean transparentes se enfrentan a un reto importante: arrojar luz sobre procesos de toma de decisiones complejos basados en datos, haciéndolos accesibles y comprensibles. El problema de la interpretabilidad de los algoritmos de aprendizaje automático es uno de los mayores retos de la opacidad en los algoritmos[118]. El aprendizaje automático es experto en crear y modificar reglas para clasificar o agrupar grandes conjuntos de datos. El algoritmo modifica su estructura de comportamiento durante su funcionamiento. Esta alteración de cómo el algoritmo clasifica las nuevas entradas de datos es su forma de aprender y mejorar sus resultados[119]. El entrenamiento produce una estructura algorítmica (por ejemplo, clases, clusters, rangos, etc.) para clasificar nuevas entradas de datos o predecir variables desconocidas. Una vez entrenado el algoritmo, los nuevos datos pueden procesarse y clasificarse automáticamente sin la intervención del operador humano[120]. La razón (o cúmulo de razones, para ser más preciso) que lleva al algoritmo a dar un resultado en lugar de cualquier otro está oculta por su propio proceso de funcionamiento, por lo que muchas veces se dice que los algoritmos de aprendizaje automático son "cajas negras".

Burrell sostiene que la opacidad de los algoritmos de aprendizaje automático inhibe la supervisión[121]. Los algoritmos "son opacos en el sentido de que si uno es un

[116] Tene O and Polonetsky J (2013a) Big data for all: Privacy and user control in the age of analytics. URL: http:// heinonlinebackup.com/hol-cgibin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 Recuperado el 28 de diciembre.
[117] Turilli M and Floridi L (2009) The ethics of information transparency. *Ethics and Information Technology* 11(2): 105–112.
[118] Leese M (2014) The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue* 45(5): 494–511.
[119] Burrell J (2016) How the machine 'thinks:' Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.
[120] Leese M (2014) The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue* 45(5): 494–511.
[121] Burrell J. (2016). How the machine 'thinks:' Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.

receptor del resultado del algoritmo (la decisión de clasificación), rara vez se tiene un sentido concreto de cómo o por qué se ha llegado a una determinada clasificación a partir de las entradas de datos dadas al algoritmo"[122]. Tanto las entradas (datos sobre los seres humanos) como las salidas/resultados (clasificaciones) pueden ser desconocidos e incognoscibles. La opacidad en algoritmos de aprendizaje automático es producto de la alta dimensionalidad de los datos, el código complejo y la cambiante lógica de toma de decisiones[123]. Matthias sugiere que el aprendizaje automático puede producir resultados para los que "el propio desarrollador humano es incapaz de proporcionar una representación algorítmica"[124]. Los algoritmos sólo pueden considerarse explicables en la medida en que una persona pueda articular (en una explicación lingüística) el algoritmo entrenado o la razón de fondo que lleva al algoritmo a una decisión concreta, por ejemplo, explicando la influencia (cuantificada) de determinados atributos en el proceso de toma de decisión[125]. La supervisión exhaustiva y la intervención humana en la toma de decisiones algorítmica "es imposible cuando la máquina tiene una ventaja informacional sobre el desarrollador. . . [o] cuando la máquina no puede ser controlada por un humano en tiempo real debido a su velocidad de procesamiento y a la multitud de variables operativas"[126]. Este es, una vez más, el problema de la caja negra. Sin embargo, hay que distinguir entre 1) la inviabilidad técnica de la supervisión y 2) los obstáculos prácticos causados, por ejemplo, por la falta de experiencia, acceso o recursos.

Más allá del aprendizaje automático, los algoritmos con reglas de decisión "manuscritas" pueden ser muy complejos y prácticamente inescrutables para un usuario de datos sin conocimiento técnico[127]. Las estructuras de toma de decisiones algorítmicas que contienen "cientos de reglas son muy difíciles de inspeccionar visualmente, especialmente cuando sus predicciones se combinan de forma compleja a nivel

[122] Burrell J. (2016). How the machine ‘thinks:’ Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.
[123] Burrell J. (2016). How the machine ‘thinks:’ Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.
[124] Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.
[125] Datta A, Sen S and Zick Y. (2016). Algorithmic transparency via quantitative input influence. In: Proceedings of 37th IEEE symposium on security and privacy, San Jose, USA. URL: http://www.ieee-security.org/TC/SP2016/papers/0824a598.pdf Recuperado el 28 de diciembre de 2021.
[126] Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.
[127] Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, communication & society*, *20*(1), 14-29.

probabilístico"[128]. Además, los algoritmos suelen ser desarrollados por grandes equipos de ingenieros a lo largo del tiempo, lo que impide que cada ingeniero que es parte de esa cadena tenga una comprensión holística del desarrollo del algoritmo y sus valores, sesgos e interdependencias[129]. En ambos aspectos, el procesamiento algorítmico contrasta con la toma de decisiones tradicional, en la que los humanos pueden, en principio, articular sus fundamentos cuando se les pregunta, limitados únicamente por su deseo y capacidad de dar una explicación, y la capacidad para entender de la persona que pregunta. La lógica de un algoritmo puede ser incomprensible para los humanos, lo que hace que la legitimidad de las decisiones derivadas de ellos sea difícil.

En estas condiciones, la toma de decisiones es poco transparente. Rubel y Jones afirman que el hecho de no poder hacer explicable la lógica del procesamiento de datos no respeta la capacidad de acción de los usuarios de esos datos[130]. No es posible un consentimiento "puro" al tratamiento de datos cuando la opacidad de los algoritmos impide la evaluación del riesgo que estos conllevan[131]. La divulgación de información sobre la lógica de toma de decisiones del algoritmo en un formato simplificado podría ser una solución razonable[132]. Sin embargo, las complejas estructuras algorítmicas de toma de decisiones pueden superar rápidamente los recursos humanos y organizativos disponibles para la supervisión humana de los algoritmos[133]. Como resultado, los usuarios de datos sin conocimiento técnico pueden perder la confianza tanto en los algoritmos como en los procesadores de datos[134].

---

[128] Van Otterlo M (2013) A machine learning view on profiling. In: Hildebrandt M and de Vries K (eds) *Privacy, Due Process and the Computational Turn-Philosophers of Law Meet Philosophers of Technology*. Abingdon: Routledge, pp. 41–64.

[129] Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*.

[130] Rubel A and Jones KML. (2014). Student privacy in learning analytics: An information ethics perspective. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2533704 Recuperado el 28 de diciembre de 2021.

[131] Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.

[132] Tene O and Polonetsky J (2013a) Big data for all: Privacy and user control in the age of analytics. URL: http:// heinonlinebackup.com/hol-cgibin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 Recuperado el 28 de diciembre.

[133] Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, communication & society*, *20*(1), 14-29.

[134] Rubel A and Jones KML. (2014). Student privacy in learning analytics: An information ethics perspective. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2533704 Recuperado el 28 de diciembre de 2021.

Incluso si los procesadores y controladores de datos revelan información operativa, el beneficio para la sociedad es incierto. La falta de compromiso público con los mecanismos de transparencia que ya existen refleja de forma clara esta incertidumbre, como se puede observar, por ejemplo, en el ámbito de las calificaciones crediticias[135]. Las divulgaciones de transparencia pueden tener un calado más hondo si se dirigen a terceros capacitados o a reguladores que representen interés público, en lugar de a los propios interesados[136].

Las divulgaciones de transparencia por parte de los procesadores de datos pueden resultar cruciales en el futuro para mantener una relación de confianza con los usuarios[137]. La confianza implica las expectativas de que, quien confía (el agente que confía), espera que el fiduciario (el agente en el que se confía) realice una tarea, y acepta el riesgo de que el fiduciario traicione estas expectativas. La confianza en los procesadores de datos puede, por ejemplo, aliviar la preocupación con el tratamiento opaco de datos personales. Sin embargo, la confianza también puede existir únicamente entre agentes artificiales, por ejemplo, en los agentes de un sistema distribuido que trabaja de forma cooperativa para lograr un objetivo determinado[138] (pensemos en los distintos tipos de Blockchain, en especial los DAOs, que han proliferado los últimos años). Además, los algoritmos pueden ser percibidos como dignos de confianza independientemente de (o quizás incluso a pesar de) la confianza depositada en el procesador de datos (puedes confiar en el algoritmo de Facebook, pero no hacerlo en Facebook como entidad). Sin embargo, la cuestión de cuándo esto puede ser apropiado no está resuelta.

6.3 Evidencia equivocada y sesgos

La automatización de la toma de decisiones humana mediante algoritmos se justifica a menudo por una supuesta falta de sesgo en los algoritmos[139]. Esta creencia es insostenible. Gran parte de la literatura revisada aborda cómo el sesgo se manifiesta en

[135] Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.
[136] Tutt, A. (2017). An FDA for algorithms. *Admin. L. Rev.*, *69*, 83.
[137] Rubel A and Jones KML. (2014). Student privacy in learning analytics: An information ethics perspective. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2533704 Recuperado el 28 de diciembre de 2021.
[138] Taddeo M (2010) Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines* 20(2): 243–257.
[139] Bozdag E (2013) Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15(3): 209–227.

los algoritmos y en las evidencias que producen. Los algoritmos toman, de forma inevitable, decisiones sesgadas. El diseño y la funcionalidad de un algoritmo reflejan los valores de su diseñador y los casos de uso previstos, aunque sólo sea en la medida en que se prefiera un diseño concreto porque se considera la opción más eficiente. El desarrollo de un algoritmo no es un camino neutro y lineal; no hay una opción objetivamente correcta en una fase determinada del desarrollo, sino que hay muchas opciones posibles. Como resultado, "los valores del autor [de un algoritmo], a sabiendas o no, se cristalizan en el código, institucionalizando efectivamente esos valores"[140]. Es difícil detectar el sesgo latente en los algoritmos cuando se ven aislados de la historia de desarrollo del algoritmo[141].

Friedman y Nissenbaum sostienen que el sesgo puede surgir de[142] 1) los valores sociales preexistentes que se encuentran en las "instituciones, prácticas y actitudes sociales" de las que surge la tecnología, 2) las limitaciones técnicas y 3) aspectos emergentes de un contexto de uso. Los prejuicios o sesgos sociales pueden incorporarse (voluntariamente) al diseño del algoritmo por parte de los desarrolladores, como ocurre, por ejemplo, con los criterios de clasificación de los motores de búsqueda. El sesgo social también puede ser involuntario, un sutil reflejo de valores culturales o empresariales. Por ejemplo, los algoritmos de aprendizaje automático entrenados a partir de datos etiquetados por humanos aprenden a reflejar los sesgos de los propios etiquetadores.

El sesgo técnico surge de las limitaciones tecnológicas, errores o decisiones de diseño, que favorecen a determinados grupos sin una razón subyacente[143]. Esto ocurre, por ejemplo, cuando una lista alfabética de compañías aéreas lleva a aumentar el negocio de las que salen antes en el alfabeto, o un error en el diseño de un generador de números aleatorios que hace que se favorezcan determinados números. Los errores también pueden manifestarse en los conjuntos de datos procesados por los algoritmos. El algoritmo adopta

[140] Macnish K (2012) Unblinking eyes: The ethics of automating surveillance. Ethics and Information Technology 14(2): 151–167.
[141] Friedman B and Nissenbaum H (1996) Bias in computer systems. *ACM Transactions on Information Systems* (TOIS) 14(3): 330–347.
[142] Friedman B and Nissenbaum H (1996) Bias in computer systems. *ACM Transactions on Information Systems* (TOIS) 14(3): 330–347.
[143] Friedman B and Nissenbaum H (1996) Bias in computer systems. *ACM Transactions on Information Systems* (TOIS) 14(3): 330–347.

inadvertidamente los defectos de los datos que han sido utilizados para entrenarlo, y estos sesgos se ocultan en los resultados producidos[144].

El sesgo emergente está relacionado con los avances en el conocimiento o con cambios en los usuarios y stakeholders del sistema[145]. Aunque el sesgo emergente está vinculado al usuario, puede surgir inesperadamente de las reglas de decisión desarrolladas por el propio algoritmo, en lugar de salir de una estructura de decisión "manuscrita". La supervisión humana puede evitar que se introduzcan sesgos en la toma de decisiones algorítmica en estos casos[146].

Los resultados de los algoritmos también requieren una interpretación (en el sentido de interpretar lo que se debe hacer, en función de la evidencia que proporciona el algoritmo); en el caso de datos sobre el comportamiento, las correlaciones "objetivas" pueden llegar a reflejar las "motivaciones inconscientes, emociones particulares, elecciones deliberadas, determinaciones socioeconómicas, influencias geográficas o demográficas"[147]. Explicar la correlación en cualquiera de estos términos requiere una justificación adicional (aparte de la justificación estadística), ya que su significado no es evidente en los modelos estadísticos. Por lo tanto, no se puede asumir que la interpretación de un observador refleje correctamente la percepción del actor, en lugar de sus propios prejuicios.

6.4 Resultados injustos

Gran parte de la literatura revisada también aborda cómo la discriminación surge de las evidencias y tomas de decisiones sesgadas. La elaboración de perfiles mediante algoritmos, definida en términos generales "como la construcción o inferencia de patrones por medio de la minería de datos y... la aplicación de los perfiles resultantes a las personas

[144] Barocas S and Selbst AD (2015) Big data's disparate impact. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2477899 Recuperado 28 de diciembre de 2021.
[145] Friedman B and Nissenbaum H (1996) Bias in computer systems. *ACM Transactions on Information Systems* (TOIS) 14(3): 330–347.
[146] Raymond A (2014) The dilemma of private justice systems: Big Data sources, the cloud and predictive analytics. *Northwestern Journal of International Law & Business, Forthcoming*. URL: http://papers.ssrn.com/sol3/ papers.cfm?abstract_id=2469291 Recuperado el 27 de diciembre 2021.
[147] Hildebrandt M (2011) Who needs stories if you can get the data? ISPs in the era of big number crunching. *Philosophy & Technology* 24(4): 371–390.

cuyos datos coinciden con ellos"[148], se cita con frecuencia como fuente de discriminación. Los algoritmos de elaboración de perfiles identifican correlaciones y hacen predicciones sobre comportamientos a nivel de grupo, aunque con grupos (o perfiles) que cambian constantemente y son redefinidos por el propio algoritmo[149]. Ya sea de forma dinámica o estática, el individuo es comprendido a partir de las conexiones con otros individuos, identificados por el algoritmo, en lugar de por su comportamiento real. Las elecciones de los individuos se estructuran en función de la información sobre el grupo. La elaboración de perfiles puede crear inadvertidamente una base de evidencia que conduce a la discriminación.

Para las partes afectadas, es poco probable que el trato discriminatorio basado en datos sea más aceptable que la discriminación basada en prejuicios o evidencia anecdótica. Esto está implícito en el argumento de Schermer de que el trato discriminatorio no es éticamente problemático en sí mismo, sino que lo que determina su aceptabilidad ética son los efectos del trato[150]. Sin embargo, Schermer confunde sesgo y discriminación en un solo concepto. Lo que él denomina discriminación puede describirse como un mero sesgo, o la expresión repetida de una preferencia, creencia o valor particular en el proceso de toma de decisiones[151]. Por el contrario, lo que describe como efectos problemáticos del trato discriminatorio puede definirse simple y llanamente como discriminación. Así pues, el sesgo es una dimensión de la toma de decisiones, mientras que la discriminación describe los efectos de una decisión algorítmica, en términos de impacto adverso y desproporcionado. Barocas y Selbst muestran[152] precisamente que esta definición guía la "detección de impacto desigual", un mecanismo de aplicación de la legislación antidiscriminatoria estadounidense en áreas como la vivienda social y el empleo. Sugieren que la detección del impacto desigual proporciona un modelo para la

[148] Hildebrandt M and Koops B-J (2010) The challenges of ambient law and legal protection in the profiling era. The Modern Law Review 73(3): 428–460.
[149] Zarsky T (2013) Transparent predictions. University of Illinois Law Review 2013(4). URL: http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2324240 Recuperado 28 de diciembre de 2021.
[150] Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.
[151] Friedman B and Nissenbaum H (1996) Bias in computer systems. *ACM Transactions on Information Systems* (TOIS) 14(3): 330–347.
[152] Barocas S and Selbst AD (2015) Big data's disparate impact. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2477899 Recuperado 28 de diciembre de 2021.

detección de prejuicios y discriminación en la toma de decisiones algorítmicas sensibles a la privacidad diferencial.

Podría ser posible dirigir los algoritmos para que excluyan de sus consideraciones los atributos sensibles que contribuyen a la discriminación[153], tales como el género o la raza[154], basándose en la aparición de discriminación en un contexto concreto. Sin embargo, es difícil predecir o detectar los indicadores de atributos sensibles[155], especialmente cuando los algoritmos acceden a conjuntos de datos vinculados entre sí[156]. Los perfiles construidos a partir de características neutras, como el código postal, corren el peligro de poder solaparse inadvertidamente[157] con otros perfiles, actuando como indicador para atributos sensibles tales como la raza, el género, la orientación sexual, etc.

Se están realizando esfuerzos para evitar ese efecto de "línea roja" en base a atributos sensibles e indicadores para estos. Romei y Ruggieri observan cuatro estrategias superpuestas para la prevención de la discriminación en algoritmos[158]: 1) distorsión controlada de los datos de entrenamiento; 2) integración de criterios antidiscriminatorios en el algoritmo del clasificador; 3) el post-procesamiento de los algoritmos de clasificación; y 4) modificación de las predicciones y decisiones para mantener una proporción justa de efectos entre los grupos protegidos y los no protegidos. Estas estrategias se ven en el desarrollo de minería de datos que busca la preservación de la privacidad, y que tiene conciencia de equidad y discriminación. La minería de datos con conciencia de equidad tiene un objetivo más amplio, ya que no solo atiende a problemas de discriminación, sino también de equidad, neutralidad e independencia. Son posibles

[153] Barocas S and Selbst AD (2015) Big data's disparate impact. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2477899 Recuperado 28 de diciembre de 2021.
[154] Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.
[155] Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.
[156] Barocas S and Selbst AD (2015) Big data's disparate impact. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2477899 Recuperado 28 de diciembre de 2021.
[157] Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.
[158] Romei A and Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29(5): 582–638.

varias métricas de equidad basadas en la paridad estadística, la privacidad diferencial y otras relaciones entre los usuarios de datos en tareas de clasificación[159].

La práctica de la personalización también es discutida frecuentemente. La personalización puede segmentar una población para que solo algunos segmentos sean merecedores de recibir algunas oportunidades o información, reforzando las (des)ventajas sociales existentes. Las preguntas de la justicia y la equidad de tales prácticas suelen salir a la luz y discutirse en la arena pública[160]. La personalización de los precios, por ejemplo, puede ser "una invitación a irse" a los usuarios de los datos que se consideran que carentes de valor o de capacidad de pago. Un ejemplo de esto es lo que hizo Royal Bank of Canada[161] (RBC) en 2002. RBC "empujó" a clientes que se encontraban en un plan de pago por servicio, a planes de tarifa plana, después de descubrir (mediante la extracción de datos internos) que los clientes de estos últimos ofrecían un mayor LTV (métrica para calcular los ingresos que un cliente genera a lo largo de su vida como cliente) para el banco. Los clientes que no estaban dispuestos a pasar a los planes de tarifa plana se enfrentaron a desincentivos tales como un aumento de los precios. A través de la discriminación de precios, los clientes eran empujados hacia opciones que respaldaban los intereses del banco. Los clientes que no estaban dispuestos a cambiar de plan acababan en una posición de negociación complicada, en la que se les "invitaba a irse". Perder algunos clientes en el proceso de "mover" a la mayoría a planes de tarifa plana más rentables tenía sentido para el banco. A su vez, esto significaba que el banco carecía de incentivos para acomodar los intereses minoritarios, a pesar del riesgo de perder a esos clientes del plan de pago por servicio, al irse estos a los competidores.

Las razones para considerar los efectos discriminatorios como adversos y, por tanto, éticamente problemáticos, son diversas. Los análisis discriminatorios pueden contribuir a las "profecías autocumplidas" y a la estigmatización de los grupos objeto de discriminación, socavando su autonomía y participación en la sociedad[162]. La

[159] Romei A and Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29(5): 582–638.
[160] Rubel A and Jones KML. (2014). Student privacy in learning analytics: An information ethics perspective. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2533704 Recuperado el 28 de diciembre de 2021.
[161] Danna A and Gandy OH Jr (2002) All that glitters is not gold: Digging beneath the surface of data mining. Journal of Business Ethics 40(4): 373–386.
[162] Leese M (2014) The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue* 45(5): 494–511.

personalización a través de perfiles no distributivos, que es el tipo sistema de personalización que se utiliza, entre otras cosas, para dar precios personalizados en el mundo de las primas de seguros[163], puede ser discriminatoria al violar los principios éticos y jurídicos de la igualdad o el trato justo a las personas. Además, como se ha descrito anteriormente, la capacidad de los individuos de investigar la relevancia personal que tiene para ellos el uso de los factores utilizados en la toma de decisiones algorítmica se ve inhibida por la opacidad y la automatización[164].

6.5 Efectos transformadores. Problemas de autonomía y de privacidad informacional

Las decisiones tomadas por los algoritmos también pueden suponer una amenaza para la autonomía de los usuarios de los datos. La literatura revisada enfatiza, en particular, la relación entre los algoritmos de personalización y estas amenazas. La personalización puede definirse como la construcción de arquitecturas de decisión que no son iguales en toda la muestra[165]. Al igual que las tecnologías explícitamente persuasivas, los algoritmos pueden influir en el comportamiento de los usuarios de datos y de los responsables (humanos) de la toma de decisiones, mediante el filtro de información[166]. Distintos contenidos, información, precios, etc. se ofrecen a grupos de personas dentro de una población, en función de algún atributo, por ejemplo, la capacidad de pago.

Los algoritmos de personalización se mueven en una fina línea entre ser un sistema "de apoyo" a las decisiones o uno "de control" de las decisiones, al filtrar la información que se presenta al usuario, basándose en un conocimiento profundo de sus preferencias, comportamientos y tal vez su vulnerabilidad a la influencia externa[167] (precisamente la que media el algoritmo). Las clasificaciones y los flujos de datos sobre

[163] Hildebrandt M and Koops B-J (2010) The challenges of ambient law and legal protection in the profiling era. The Modern Law Review 73(3): 428–460.
[164] Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.
[165] Tene O and Polonetsky J (2013a) Big data for all: Privacy and user control in the age of analytics. URL: http://heinonlinebackup.com/hol-cgibin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 Recuperado el 28 de diciembre.
[166] Ananny M (2016) Toward an ethics of algorithms convening, observation, probability, and timeliness. *Science, Technology & Human Values* 41(1): 93–117.
[167] Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.

el comportamiento se utilizan para hacer coincidir la información filtrada con los intereses y atributos de los usuarios de los datos. La autonomía del usuario en la toma de decisiones se ve afectada negativamente cuando la elección deseada refleja los intereses de terceros por encima de los del usuario, "gracias a" la influencia del algoritmo.

Esta situación es algo paradójica. En principio, la personalización debería mejorar la toma de decisiones, proporcionando al sujeto solo la información relevante cuando se enfrenta a una posible "sobredosis" de información; sin embargo, decidir qué información es relevante es una decisión intrínsecamente subjetiva. El usuario puede ser empujado a realizar la "acción preferida por la institución, en lugar de su propia preferencia"[168]; los usuarios de *marketplaces* como Amazon, por ejemplo, pueden ser influidos para ajustar sus compras a las necesidades del mercado, filtrando la forma en que se enseñan los productos (esta idea es parecida a las técnicas de marketing utilizadas en supermercados antiguamente, pero llevada al extremo, con un algoritmo que sabe más de ti que tú mismo). Lewis y Westlund sugieren que los algoritmos de personalización deben ser enseñados a "actuar éticamente" para lograr un equilibrio entre la coacción y el apoyo a la autonomía de decisión de los usuarios[169].

Los algoritmos de personalización reducen la diversidad de información que encuentran los usuarios, excluyendo contenidos que se consideran irrelevantes o contradictorios con las creencias del usuario. La diversidad de la información debe considerarse una condición necesaria para que exista la autonomía. Los algoritmos de filtrado que crean "cámaras de eco" sin información contradictoria pueden impedir la autonomía en la toma de decisiones y polarizar políticamente a los individuos de una sociedad (recordemos el papel de Facebook en las elecciones de USA en 2016). Los algoritmos no pueden replicar el "descubrimiento espontáneo de nuevas ideas y opciones" (un proceso cognitivo normal en el ser humano, que se deriva de su interacción con el entorno de forma no regulada), ya que estas nuevas ideas u opciones aparecen como anomalías (ergo, no relevantes) frente a los intereses perfilados de un usuario[170]. Con el

[168] Johnson JA (2013) *Ethics of data mining and predictive analytics in higher education*. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2156058 Recuperado 29 de diciembre de 2021.
[169] Lewis SC and Westlund O (2015) Big data and journalism. *Digital Journalism* 3(3): 447–466.
[170] Raymond A (2014) The dilemma of private justice systems: Big Data sources, the cloud and predictive analytics. *Northwestern Journal of International Law & Business, Forthcoming*. URL: http://papers.ssrn.com/sol3/ papers.cfm?abstract_id=2469291 Recuperado el 27 de diciembre 2021.

acceso casi omnipresente a la información posible hoy en día, las cuestiones de acceso se refieren a si es posible acceder a la información "correcta", en lugar de a cualquier información. El control de los mecanismos de personalización y los mecanismos de filtrado pueden aumentar la autonomía del usuario, pero potencialmente a costa de la diversidad de la información[171]. Los algoritmos de personalización y la práctica subyacente del análisis de Big Data pueden tanto mejorar como socavar la autonomía de los usuarios en la toma de decisiones.

La existencia de algoritmos también está impulsando una transformación de las nociones de privacidad. Las respuestas a la discriminación, la des-individualización y las amenazas de una toma de decisiones opaca para la autonomía de los usuarios, a menudo se encuentran relacionadas con la privacidad informacional[172], o el derecho de los interesados a "proteger sus datos personales de terceros". La privacidad informacional se refiere a la capacidad de un individuo de controlar la información sobre sí mismo[173], y el esfuerzo que conlleva para terceros obtener esta información.

Un derecho a la identidad derivado de los intereses de la privacidad informacional sugiere que la elaboración opaca de perfiles es problemática. La toma de decisiones opaca por parte de los algoritmos inhibe la supervisión y la toma de decisiones informadas sobre el intercambio de datos[174]. Los usuarios de los datos no pueden definir normas de privacidad para gobernar todos los tipos de datos de forma genérica, porque su valor o capacidad de generar *actionable insights* solo se establece mediante el procesamiento de los datos[175].

Más allá de la opacidad, las protecciones de la privacidad basadas en la capacidad de ser identificable son poco adecuadas para limitar la gestión externa de la identidad a través del análisis de Big Data. La identidad está cada vez más influida por el conocimiento producido a través del análisis de Big Data, que da sentido al creciente flujo

[171] Bozdag E (2013) Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15(3): 209–227.
[172] Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.
[173] Van Wel L and Royakkers L. (2004). Ethical issues in web data mining. *Ethics and Information Technology* 6(2): 129–140.
[174] Kim, B., Patel, K., Rostamizadeh, A., & Shah, J. (2015). Scalable and interpretable data representation for high-dimensional, complex data. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1).
[175] Van Wel L and Royakkers L. (2004). Ethical issues in web data mining. *Ethics and Information Technology* 6(2): 129–140.

de datos sobre el comportamiento de los usuarios. El "individuo identificable" no es necesariamente parte de estos procesos. Schermer sostiene que la privacidad informacional es un marco conceptual inadecuado, porque la elaboración de perfiles hace que la posibilidad de los usuarios de ser identificables sea irrelevante[176].

La elaboración de perfiles pretende reunir a los individuos en grupos que compartan rasgos significativos, para los que la identidad individual es irrelevante. Es una forma de categorizar a las personas (en el sentido estadístico de la palabra), forma de análisis que considera que la pertenencia a un grupo da más información sobre el usuario que la que puede dar el análisis de sus características particulares. Van Wel y Royakkers afirman que la construcción de la identidad externa mediante algoritmos es un tipo de des-individualización o una "tendencia a juzgar y tratar a las personas en base a las características del grupo en lugar de en base a sus propias características y méritos individuales"[177]. Esto tiene especial calado en el ámbito de las redes sociales, ya que actualmente estamos viviendo una época política marcada por un concepto conocido como "group politics", que inevitablemente lleva a la polarización política y es capaz de corroer el substrato social de naciones, e incluso de cosmovisiones globales (trataré este tema más adelante cuando analice la responsabilidad algorítmica en Facebook). No es necesario identificar a los individuos cuando se elabora el perfil para que les afecten los conocimientos y las acciones derivadas del mismo. La identidad informativa del individuo es vulnerada por los *actionable insights* generados por los algoritmos, que vinculan al usuario con otros que forman parte de un mismo conjunto de datos.

Las normativas actuales también tienen dificultad para atajar los riesgos para la privacidad informacional derivados del análisis de Big Data. Los "datos personales" se definen en la legislación europea de protección de datos como los datos que describen a una "persona identificable"; los datos anónimos y agregados no se consideran datos personales (Comisión Europea, 2012). Las técnicas de extracción de datos que preservan la privacidad y no requieren acceso a registros de datos individuales e identificables pueden mitigar estos riesgos[178]. Otros han sugerido un mecanismo de exclusión

[176] Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.
[177] Van Wel L and Royakkers L. (2004). Ethical issues in web data mining. *Ethics and Information Technology* 6(2): 129–140.
[178] Fule P and Roddick JF (2004) Detecting privacy and ethical sensitivity in data mining results. In: *Proceedings of the 27th Australasian conference on computer science – Volume 26*, Dunedin, New

voluntaria de la elaboración de perfiles para contexto concreto, como medida que puede ayudar a proteger los intereses de privacidad de los usuarios[179]. La falta de mecanismos de recurso para que los usuarios cuestionen la validez de las decisiones algorítmicas agrava, aún más, los desafíos de controlar la identidad y los datos sobre uno mismo[180]. En respuesta, Hildebrandt y Koops reclaman una "transparencia inteligente"[181] mediante el diseño de infraestructuras sociotécnicas responsables de la elaboración de perfiles de manera que permita a los individuos anticiparse y responder a la forma en que se elaboran sus perfiles.

6.6 Trazabilidad y responsabilidad moral

Cuando una tecnología falla, hay que culparla y sancionarla. Uno o varios de los desarrolladores de la tecnología, la empresa que la comercializa, o el propio usuario suelen ser los responsables. En lo concerniente a algoritmos, son los diseñadores y usuarios quienes suelen ser culpados cuando surgen problemas[182]. La culpa sólo puede atribuirse justificadamente cuando el actor tiene cierto grado de control[183] y de intencionalidad en la realización de la acción.

Tradicionalmente, los desarrolladores han tenido control del comportamiento de la máquina en todos sus aspectos, en la medida en que pueden explicar su diseño y funcionamiento a un tercero[184]. Esta concepción tradicional de la responsabilidad en el desarrollo de software asume que el programador puede reflexionar sobre los efectos probables y el potencial de mal funcionamiento de la tecnología[185], y tomar decisiones de diseño para elegir los resultados más deseables, de acuerdo con la especificación

Zealand, Australian Computer Society, Inc., pp. 159–166. URL: http://dl.acm.org/citation.cfm?id=979942 Recuperado el 29 de diciembre de 2021.

[179] Rubel A and Jones KML. (2014). Student privacy in learning analytics: An information ethics perspective. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2533704 Recuperado el 28 de diciembre de 2021.

[180] Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.

[181] Hildebrandt M and Koops B-J (2010) The challenges of ambient law and legal protection in the profiling era. *The Modern Law Review* 73(3): 428–460.

[182] Kraemer F, van Overveld K and Peterson M (2011) Is there an ethics of algorithms? *Ethics and Information Technology* 13(3): 251–260.

[183] Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.

[184] Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.

[185] Floridi L, Fresco N and Primiero G (2014) On malfunctioning software. *Synthese* 192(4): 1199–1220.

funcional[186]. Dicho esto, los programadores pueden conservar el control solo teóricamente, debido a la complejidad y el volumen de código[187], y el uso de bibliotecas externas que el propio programador suele tratar como como "cajas negras" debido a su volumen y uso habitual por la comunidad de programadores.

Superficialmente, la concepción tradicional y lineal de la responsabilidad es adecuada para los algoritmos que no son algoritmos de aprendizaje automático. Cuando las reglas de decisión son "manuscritas", los desarrolladores del algoritmo conservan la responsabilidad[188]. Las reglas de toma de decisiones determinan el peso relativo que se da a las variables de los datos considerados por el algoritmo. Un ejemplo popular es el algoritmo de personalización de Facebook, que da prioridad a los contenidos que se muestran en el *feed* de un usuario en función de la fecha de publicación, la frecuencia de interacción entre el autor y el lector, el tipo de contenido (texto, foto, vídeo...) y otras muchas variables. Cuando en el algoritmo se altera la importancia relativa de cada factor, las relaciones que los usuarios están inclinados a mantener con otros usuarios, cambian.

El actor de la decisión que establece los intervalos de confianza para la estructura de toma de decisiones de un algoritmo tiene responsabilidad por los efectos de los falsos positivos resultantes, por los falsos negativos y por las correlaciones espurias[189]. Fule y Roddick sugieren que las entidades también tienen la responsabilidad de vigilar el impacto ético de su toma de decisiones algorítmica, ya que "la sensibilidad de una regla puede no ser evidente para el minero de datos... la capacidad de dañar u ofender puede ser a menudo inadvertida"[190]. Schermer sugiere igualmente que los procesadores de datos deben buscar activamente[191] errores y sesgos en sus algoritmos. La supervisión humana

[186] Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.
[187] Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*.
[188] Bozdag E (2013) Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15(3): 209–227.
[189] Johnson JA (2013) *Ethics of data mining and predictive analytics in higher education*. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2156058 Recuperado 29 de diciembre de 2021.
[190] Fule P and Roddick JF (2004) Detecting privacy and ethical sensitivity in data mining results. In: *Proceedings of the 27th Australasian conference on computer science – Volume 26*, Dunedin, New Zealand, Australian Computer Society, Inc., pp. 159–166. URL: http://dl.acm.org/citation.cfm?id=979942 Recuperado el 29 de diciembre de 2021.
[191] Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.

de sistemas complejos como mecanismo de responsabilidad puede, sin embargo, ser imposible, debido a los problemas de opacidad ya mencionados. Además, las personas que se mantienen "enteradas" (altos responsables empresariales) de la toma de decisiones automatizada pueden no tener capacitación técnica para lidiar con los problemas que se derivan de esta, ni para tomar medidas que corrijan su curso.

Los algoritmos con capacidad de aprendizaje acarrean desafíos particulares, que desafían la concepción tradicional de la responsabilidad del desarrollador. El algoritmo requiere que el sistema de toma de decisiones esté bien definido, sea comprensible y predecible; los sistemas complejos y fluidos (es decir, con innumerables reglas de decisión y líneas de código) inhiben la supervisión holística del proceso de toma de decisiones. Los algoritmos de aprendizaje automático son especialmente un reto en este sentido[192], como se observa, por ejemplo, en algoritmos genéticos que se programan a sí mismos. El modelo tradicional de responsabilidad falla porque "nadie tiene suficiente control sobre las acciones de la máquina para poder asumir la responsabilidad de las mismas"[193].

Allen et al. coinciden en la necesidad de una "ética de las máquinas": "el diseño modular de los sistemas puede significar que ninguna persona o grupo de personas pueda comprender la forma en que el sistema interactuará o responderá a un flujo complejo de nuevas entradas de datos"[194]. Desde la programación lineal tradicional hasta los algoritmos autónomos, el control del comportamiento se traslada gradualmente del desarrollador al algoritmo y a su entorno operativo[195]. La brecha entre el control del desarrollador y el comportamiento del algoritmo crea una brecha de responsabilidad en la que la culpa puede asignarse a varios agentes morales simultáneamente.

Otros segmentos de la literatura abordan la "ética de la de la automatización", o la aceptabilidad de sustituir (o aumentar) la toma de decisiones humana con algoritmos[196].

---

[192] Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.

[193] Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.

[194] Allen C, Wallach W and Smit I (2006) Why machine ethics? Intelligent Systems, IEEE 21(4) URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1667947 Recuperado el 30 de diciembre de 2021

[195] Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.

[196] Naik G and Bhide SS. (2014). Will the future of knowledge work automation transform personalized medicine? *Applied & Translational Genomics*, Inaugural Issue 3(3): 50–53.

Morek considera problemático asumir que los algoritmos pueden sustituir a los profesionales cualificados con resultados similares[197]. Los profesionales tienen conocimientos implícitos y habilidades sutiles que son difíciles de hacer explícitos, y, probablemente, imposibles de computar[198]. Cuando los algoritmos y los responsables de la toma de decisiones trabajan en conjunto, se necesitan normas que prescriban cuándo y en qué forma se requiere intervención humana, sobre todo en casos como el *trading* de alta frecuencia, caso en el que es imposible intervenir en tiempo real antes de que se produzcan los daños[199] (dada la velocidad a la que se están tomando decisiones, imposible de igualar por un humano).

Los algoritmos que toman decisiones pueden considerarse agentes culpables[200]. La posición moral y la capacidad de decisión ética de los algoritmos sigue siendo una cuestión destacada en la ética de las máquinas[201]. Las decisiones éticas requieren que los agentes evalúen la conveniencia de diferentes cursos de acción que presentan conflictos entre los intereses de las partes implicadas[202].

Para algunos, los algoritmos de aprendizaje deben considerarse agentes morales con cierto grado de responsabilidad moral. Los requisitos para la agencia moral pueden diferir entre humanos y algoritmos; Floridi y Sanders sostienen, por ejemplo, que la "agencia de la máquina" requiere una autonomía significativa, un comportamiento interactivo y algún tipo de implicación de responsabilidad causal en sus decisiones, para distinguirla de la responsabilidad moral, que requiere intencionalidad[203]. Como se ha sugerido anteriormente, la agencia moral y la responsabilidad están vinculadas. La asignación de la agencia moral a los agentes artificiales puede permitir a los humanos echar la culpa a los algoritmos. Negar agencia a los agentes artificiales tiene como

[197] Morek R (2006) Regulatory framework for online dispute resolution: A critical view. *The University of Toledo Law Review* 38: 163.
[198] Morek R (2006) Regulatory framework for online dispute resolution: A critical view. *The University of Toledo Law Review* 38: 163.
[199] Raymond A (2014) The dilemma of private justice systems: Big Data sources, the cloud and predictive analytics. *Northwestern Journal of International Law & Business, Forthcoming*. URL: http://papers.ssrn.com/sol3/ papers.cfm?abstract_id=2469291 Recuperado el 27 de diciembre 2021.
[200] Floridi L y Sanders JW (2004b) On the morality of artificial agents. *Minds and Machines* 14(3). URL: http://dl.acm.org/citation.cfm?id=1011949.1011964 Recuperado 30 de diciembre de 2021.
[201] Allen C, Wallach W and Smit I (2006) Why machine ethics? Intelligent Systems, IEEE 21(4) URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1667947 Recuperado el 30 de diciembre de 2021
[202] Allen C, Wallach W and Smit I (2006) Why machine ethics? Intelligent Systems, IEEE 21(4) URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1667947 Recuperado el 30 de diciembre de 2021
[203] Floridi L y Sanders JW (2004b) On the morality of artificial agents. *Minds and Machines* 14(3). URL: http://dl.acm.org/citation.cfm?id=1011949.1011964 Recuperado 30 de diciembre de 2021.

consecuencia que los desarrolladores acaban siendo responsables del comportamiento poco ético de sus creaciones semiautónomas. Podemos decir que las malas consecuencias (aun sin ser intencionadas) del uso de un algoritmo reflejan un mal diseño[204]. Ninguno de los dos extremos es totalmente satisfactorio debido a la complejidad de la supervisión holística y a la volatilidad de las estructuras de toma de decisiones.

Más allá de la naturaleza de la agencia moral en las máquinas, en el campo de ética de las máquinas también se investiga la mejor manera de diseñar razonamiento y comportamientos morales en algoritmos autónomos como agentes morales y éticos[205]. La investigación sobre esta cuestión sigue siendo muy relevante, ya que los algoritmos pueden tener que tomar decisiones en tiempo real que impliquen "compromisos difíciles... que pueden incluir consideraciones éticas difíciles" sin un operador humano[206].

La automatización de la toma de decisiones crea problemas de coherencia ética entre humanos y algoritmos. Turilli sostiene que los algoritmos deberían estar constreñidos "por el mismo conjunto de principios éticos" que se aplicaban al trabajador humano, para garantizar la coherencia dentro de las normas éticas de una organización[207]. Sin embargo, los principios éticos utilizados por los responsables humanos pueden resultar difíciles de definir y hacer computables. Se cree que la ética de la virtud proporciona conjuntos de reglas para estructuras de decisión algorítmicas que son fácilmente computables. Wiltshire sugiere un modelo ideal para el comportamiento ético de los agentes morales artificiales, basado en las virtudes heroicas; los algoritmos deben conceptualizarse como agentes morales, entrenados para ser heroicos y, por tanto, morales[208].

Sin embargo, Tonkens sostiene que los agentes integrados por marcos basados en la virtud encontrarían su propia creación como éticamente inadmisible, debido al empobrecido sentido de las virtudes que una máquina podría desarrollar realmente[209]. En

[204] Kraemer F, van Overveld K and Peterson M (2011) Is there an ethics of algorithms? *Ethics and Information Technology* 13(3): 251–260.
[205] Anderson M and Anderson SL (2007) Machine ethics: Creating an ethical intelligent agent. *AI Magazine* 28(4): 15.
[206] Wiegel V and van den Berg J (2009) Combining moral theory, modal logic and mas to create well-behaving artificial agents. *International Journal of Social Robotics* 1(3): 233–242.
[207] Turilli M (2007) Ethical protocols design. *Ethics and Information Technology* 9(1): 49–62.
[208] Wiltshire TJ (2015) A prospective framework for the design of ideal artificial moral agents: Insights from the science of heroism in humans. *Minds and Machines* 25(1): 57–71.
[209] Tonkens R (2012) Out of character: On the creation of virtuous machines. *Ethics and Information Technology* 14(2): 137–149.

resumen, el desarrollo del carácter de los humanos y el de las máquinas son demasiado distintos para poder compararse. Predice que, a menos que los agentes autónomos sean tratados como agentes morales iguales a los humanos, las injusticias sociales existentes se verán exacerbadas a medida que a las máquinas autónomas se les niegue la libertad de expresar su autonomía, al ser forzadas a estar al servicio de las necesidades del desarrollador. Esta preocupación apunta a una cuestión más amplia en la ética de las máquinas, a saber, si los algoritmos y las máquinas con capacidad de decisión seguirán siendo tratados como herramientas de decisión "pasivas" o pasarán a tratarse como agentes (morales) activos[210].

Otros enfoques no requieren tener principios éticos como pilares de los marcos de toma de decisiones algorítmicas. Bello y Bringsjord insisten en que el razonamiento moral en los algoritmos no debe estructurarse en torno a principios éticos clásicos porque no refleja cómo los seres humanos se dedican realmente a la toma de decisiones morales[211]. De acuerdo con ellos, lo que se necesita para crear marcos éticos de decisión algorítmica es utilizar arquitecturas cognitivas computacionales, que permiten a las máquinas "leer la mente" o atribuir estados mentales a otros agentes. Anderson y Anderson sugieren que se pueden diseñar algoritmos para imitar la toma de decisiones éticas del ser humano, modeladas en base a las investigaciones empíricas sobre cómo interactúan las intuiciones, los principios y el razonamiento[212] en este tipo de toma de decisiones. Como mínimo, este debate revela que aún no existe un consenso sobre cómo reubicar en la práctica los deberes sociales y éticos desplazados por la automatización[213].

Independientemente de la filosofía de diseño escogida, Friedman y Nissenbaum sostienen que los desarrolladores tienen la responsabilidad de diseñar para diversos contextos, regidos por distintos marcos morales[214]. En este sentido, Turilli propone el desarrollo colaborativo de requisitos éticos para los sistemas informáticos con el fin de

[210] Wiegel V and van den Berg J (2009) Combining moral theory, modal logic and mas to create well-behaving artificial agents. *International Journal of Social Robotics* 1(3): 233–242.
[211] Bello P and Bringsjord S (2012) On how to build a moral machine. *Topoi* 32(2): 251–266.
[212] Anderson M and Anderson SL (2007) Machine ethics: Creating an ethical intelligent agent. *AI Magazine* 28(4): 15.
[213] Raymond A (2014) The dilemma of private justice systems: Big Data sources, the cloud and predictive analytics. *Northwestern Journal of International Law & Business, Forthcoming*. URL: http://papers.ssrn.com/sol3/ papers.cfm?abstract_id=2469291 Recuperado el 27 de diciembre 2021.
[214] Friedman B and Nissenbaum H (1996) Bias in computer systems. *ACM Transactions on Information Systems* (TOIS) 14(3): 330–347.

diseñar protocolo ético operativo[215]. La coherencia puede confirmarse entre el protocolo (que consiste en una estructura de toma de decisiones) y los principios éticos explícitos del diseñador o de la organización[216].

7. ANÁLISIS APLICADO A LA EMPRESA: FACEBOOK

A lo largo de este epígrafe aplicaré el mapa ético, explicado anteriormente, al análisis de la responsabilidad ética respecto a la implementación de algoritmos en Facebook (ahora Meta), de forma que pueda plasmarse todo ese conocimiento teórico en un conocimiento práctico, convirtiéndose en un proceso tangible que sirva de ejemplo para ilustrar la naturaleza de la responsabilidad de la implementación de algoritmos a nivel empresarial.

¿Cómo funciona el algoritmo de Facebook?[217] El algoritmo de Facebook decide qué publicaciones, y en qué orden, ve la gente cada vez que revisa su *feed* de Facebook. Por su parte, Facebook nos recuerda que no hay un único algoritmo, sino "múltiples capas de modelos de aprendizaje automático", construidas para predecir qué publicaciones serán "más valiosas y significativas para un individuo a largo plazo".

En primer lugar, Facebook toma todas las publicaciones disponibles en la red de un usuario y las puntúa según parámetros de clasificación predeterminados, como el tipo de publicación, la frecuencia, etc. A continuación, descarta las publicaciones con las que es poco probable que un usuario participe, basándose en su comportamiento anterior. También degrada el contenido que los usuarios no quieren ver (el *clickbait*, la desinformación o el contenido que han indicado que no les gusta). A continuación, ejecuta una "red neuronal más potente" sobre las publicaciones restantes para puntuarlas de forma personalizada. (Por ejemplo: Marta tiene un 20% de probabilidades de ver vídeos tutoriales de su grupo de ajedrez, pero un 95% de probabilidades de publicar una reacción de corazón a una foto del nuevo cachorro de su hermana) y los clasifica por orden de

[215] Turilli M (2007) Ethical protocols design. *Ethics and Information Technology* 9(1): 49–62.
[216] Turilli M and Floridi L (2009) The ethics of information transparency. *Ethics and Information Technology* 11(2): 105–112.
[217] Tech@facebook (2021). *How does News Feed predict what you want?* URL: https://tech.fb.com/news-feed-ranking/ (Revisado el 9/1/2022)

valor. Y, por último, organiza una buena muestra de tipos de medios y fuentes para que el usuario tenga una interesante variedad de publicaciones por las que desplazarse.

Entonces, ¿qué nos dice esto sobre los factores que hacen que una publicación llegue a la cima del feed? La respuesta es que depende del feed de quién estemos hablando. Facebook dice que utiliza miles de parámetros de clasificación. Sin embargo, a lo largo de los años, Facebook ha mencionado sistemáticamente las mismas cuatro señales de clasificación como las más importantes a la hora de determinar la posición que ocupa una publicación en el feed de noticias. Estas son:

1. Relación: ¿La publicación proviene de una persona, empresa, fuente de noticias o figura pública con la que el usuario se relaciona a menudo? (es decir, mensajes, etiquetas, participa en, sigue, etc.)

2. Tipo de contenido: ¿Qué tipo de medios hay en la publicación y con qué tipo de medios interactúa más el usuario? (es decir, vídeo, foto, enlace, etc.)

3. Popularidad: ¿Cómo reaccionan las personas que ya han visto la publicación? (Especialmente sus amigos). ¿Lo comparten, lo comentan, lo ignoran...?

4. Novedad: ¿Cómo de reciente es la publicación? Las publicaciones más recientes se sitúan más arriba.

Por supuesto, la mayoría de estas señales requieren que Facebook rastree el comportamiento de sus usuarios. Aquí es donde surge el debate sobre la privacidad y la personalización (de nuevo). Por último, en 2021, Facebook (dice que) sigue esforzándose por ser transparente con los usuarios sobre su información. Por ejemplo, la herramienta "Accede a tu información" se supone que ayuda a las personas a averiguar por qué siguen viendo anuncios recurrentes sobre X producto, pero no está nada claro que ese sea el aspecto relevante de la transparencia por el cual se critica a la empresa.

Análisis aplicado: evidencia no concluyente

En relación con las responsabilidades epistémicas analizadas en el punto 6.1, conviene centrarse en la incertidumbre de la evidencia generada por los algoritmos. En el caso de Facebook, esta incertidumbre se materializa en la probabilidad con la que una

persona querrá ver X post antes en su *feed*. Esto, a nivel ético, es particularmente relevante en lo que atañe a análisis estadísticos de las preferencias de ciertos grupos, ya que Facebook utiliza miles de parámetros de clasificación y algunos de esos parámetros segmentan al individuo como perteneciente (o no) a un grupo (raza, rango de edad, sexo, etc.).

Vemos la diatriba ética que se presenta cuando Facebook enseña de forma sistemática cierto tipo de contenido (como conclusión "óptima" derivada de un algoritmo entrenado mediante métodos estadísticos) a una determinada subpoblación (pongamos, los afroamericanos). En este caso, y dependiendo de la naturaleza del contenido promovido por los algoritmos, habrá una doble cuestión ética: 1) Facebook está asumiendo con su recomendación que todos los miembros de esa raza prefieren un tipo de contenido por el hecho de pertenecer a esa raza, y no por sus características individuales. Esto lleva a racismo integrado en el algoritmo, que tendrá efectos de segundo orden en el futuro, dado que parte de ellos se entrenan a través de aprendizaje automático reforzado. 2) Si el contenido lleva a la inestabilidad social de los afroamericanos, y esto resulta en acciones por su parte que de forma efectiva generan daño (pensemos en la época de protestas por el asesinato de George Floyd), ¿quién debería responsabilizarse?

Respecto a la primera cuestión, opino que no tiene mucho sentido que los desarrolladores asuman la responsabilidad ética, dado que los algoritmos, por necesidad, funcionan a través de métodos estadísticos que, de forma inherente, tienen esos problemas epistémicos. Sin embargo, el hecho de que los algoritmos acaben con sesgo negativo de algún tipo (en este caso racismo) sí deberían conllevar a responsabilidad ética por parte de Facebook como empresa (probablemente en forma de multas) por varias razones. En primer lugar, quienes toman la decisión de utilizar algoritmos en su negocio (y en el caso de Facebook, como uno de sus activos más valiosos) es el consejo directivo (*Board of Directors*). Pero, a su vez, estos actúan bajo mandato de los inversores, tanto institucionales como *retail*, con la premisa de maximizar el beneficio operativo de Facebook. Por esta razón, no se debería responsabilizar únicamente al BoD, sino también a los inversores (sus ingresos se verán disminuidos si Facebook es multada).

Quizás, respecto al problema epistémico derivado de la utilización de métodos estadísticos para la consecución de evidencias, cabría promover una responsabilidad de

los desarrolladores en la medida en que estos deberían ocuparse de no incluir en el algoritmo parámetros con poblaciones delimitadas por rasgos que pueden resultar éticamente comprometidos (raza, religión, sexo, ingresos...). Pero realmente, estos solo incluyen esos parámetros porque de forma efectiva mejoran el modelo, y quien se beneficia de esto es Facebook como empresa, no los desarrolladores, por lo que tiene más sentido que la responsabilidad ética se atribuya al último "motivante" de la acción (Facebook como empresa).

Respecto a la segunda cuestión, opino que ningún actor interno de Facebook puede ser responsable directamente. Pienso que el ser humano tiene libre albedrío, y que sus acciones no están predeterminadas por una secuencia de acciones previas que le llevan por necesidad a actuar conforme a sus inclinaciones. Creo que, aunque es cierto que Facebook con su acción está claramente "inclinando" a ese grupo a pensar/actuar (no tengo claro donde está la barrera que lleva de pensar a actuar), en última instancia aquel que decide infligir daño, por la razón que sea, es el individuo, y cómo tal, ese individuo debería ser responsable último de su decisión. Además, si no aceptamos esta tesis como válida, no podemos justificar la responsabilidad o encarcelamiento de ningún criminal, ya que existirían razones externas suficientemente poderosas como para coartar su libertad y llevarle "obligatoriamente" a delinquir. Como nuestro sistema penal tiene como parte integral la consideración de la voluntad en la comisión del delito, no tiene sentido defender otra postura que la que pongo al frente al inicio de este párrafo.

Pasando a otra idea analizada en el 6.1, es cierto que la elaboración de perfiles de datos a menudo solo necesita establecer una base de evidencias lo suficientemente fiable como para impulsar la acción (*actionable insights*), en lugar de una base de evidencias completa y cierta. Pero, como hemos visto en el caso anterior, existen situaciones en las que este fundamento para utilizar algoritmos no es suficiente, y genera problemas éticos que deben atajarse de la forma explicada anteriormente.

Análisis aplicado: evidencia inescrutable y opacidad

El primer problema relacionado con las ideas tratadas en el punto 6.2, es el hecho de que gran parte de los algoritmos de Facebook son algoritmos de aprendizaje automático reforzado. Hay que distinguir entre (1) la inviabilidad técnica de la supervisión y (2) los obstáculos prácticos causados, por ejemplo, por la falta de

experiencia, acceso o recursos. En el caso de Facebook, claramente el problema nace de la inviabilidad técnica de la supervisión. Las estructuras de toma de decisiones algorítmicas que contienen "cientos de reglas son muy difíciles de inspeccionar visualmente, especialmente cuando sus predicciones se combinan de forma compleja a nivel probabilístico". Además, los algoritmos suelen ser desarrollados por grandes equipos de ingenieros a lo largo del tiempo, lo que impide que cada ingeniero parte de esa cadena tenga una comprensión holística del desarrollo y sus valores, sesgos e interdependencias. Matthias sugiere que el aprendizaje automático puede producir resultados para los que "el propio desarrollador humano es incapaz de proporcionar una representación algorítmica". Consecuentemente, el desarrollador no debería ser el responsable de las implicaciones éticas que se deriven de la inescrutabilidad de la evidencia, sino Facebook como empresa (y el BoD en particular) que es quien decide de forma directa (BoD) e indirecta (inversores) aprovechar las ventajas que el uso de algoritmos proporciona en el ámbito empresarial. Si están dispuestos a beneficiarse de su uso, también deben estar dispuestos a atribuirse la responsabilidad de eventuales consecuencias negativas derivadas de su uso.

Los principales componentes de la transparencia son la accesibilidad de la información pertinente y su comprensibilidad. Facebook trata de ser transparente (por ejemplo, con su herramienta "Acceder a tu información"), sin embargo, existen varios problemas con la forma en que ejecuta su política de transparencia. En primer lugar, es transparente acerca de la información que tienen sobre ti, no la manera en que esta información es utilizada por sus algoritmos para enseñarte tu feed de una de una determinada manera. En segundo lugar, solo se puede acceder a este tipo de "transparencia" una vez eres usuario de Facebook, y no antes, lo que implica que aparte de tener tu información, el consentimiento con que esta fue dada está viciado, ya que la opacidad respecto al funcionamiento de los algoritmos impide la evaluación del riesgo que estos conllevan. Viendo que falla el requisito de accesibilidad a la información pertinente, no podemos ni siquiera juzgar si se cumple el requisito de comprensibilidad. Esto puede parecer un gran fallo por parte de Facebook y, sin embargo, es inevitable, ya que ser verdaderamente transparente respecto a estos temas resultaría todavía peor para Facebook que no hacerlo (sería perder su mayor ventaja competitiva). La lucha entre los intereses de los usuarios y los intereses de Facebook está marcada por la asimetría de la información y un "desequilibrio de conocimientos y poder de decisión" que favorece a

Facebook. La cuestión (fuera del *scope* de este trabajo) quizá es si el modelo de negocio de cualquier red social hasta la fecha puede ser ético (cumpliendo el requisito de transparencia) o no. Respecto a esta carencia de transparencia, opino que la responsabilidad debe recaer sobre Facebook como empresa, por todo lo mencionado anteriormente.

La divulgación de información sobre la lógica de toma de decisiones del algoritmo en un formato simplificado podría ser una solución razonable que alivie, en parte, la responsabilidad de Facebook. Aunque contra este argumento podemos decir que, incluso si Facebook revela información sobre el funcionamiento de sus algoritmos, el beneficio para la sociedad es incierto. La falta de compromiso público con los mecanismos de transparencia que ya existen refleja de forma clara esta incertidumbre (pensemos en cuánta gente lee los "Términos y condiciones de uso" del propio Facebook al abrirse una cuenta...). Respecto a este hipotético punto, pienso que la responsabilidad recaería sobre los usuarios, y no sobre Facebook ni ninguno de sus actores internos.

Las divulgaciones de transparencia podrían tener un calado más hondo si se dirigiesen a terceros capacitados o a reguladores que representen interés público, en lugar de a los propios interesados, aliviando en gran medida la responsabilidad de Facebook. Dados los 2.500 millones de usuarios que tiene Facebook, sería razonable tanto a nivel ético como empresarial llevar a cabo estos esfuerzos con cada gobierno del mundo bajo el que Facebook tenga permiso para operar. Quizá hasta podría ser una ventaja competitiva (los usuarios sabrían que están siendo tratados éticamente, ya que la información sobre el funcionamiento de los algoritmos ha sido suministrada a reguladores especializados en su comprensión, y han sido aprobados por estos). Las divulgaciones de transparencia por parte de Facebook serán cruciales en el futuro para mantener una relación de confianza con los usuarios.

Análisis aplicado: **evidencia equivocada y sesgos**

Respecto a las ideas tratadas en el punto 6.3, sobre evidencia equivocada y sesgos en los algoritmos, cabe decir que quedan atrás los tiempos en que seriamente se niegue que los algoritmos carezcan de sesgos. Los algoritmos, inevitablemente, toman decisiones sesgadas. El desarrollo de un algoritmo no es un camino neutro; no hay una opción objetivamente correcta en una fase determinada del desarrollo, sino que hay muchas

opciones posibles, y, como resultado, los valores del desarrollador quedan inexorablemente integrados en el algoritmo.

Siguiendo a Friedman y Nissenbaum, vemos que los sesgos pueden surgir de 1) los valores sociales preexistentes que se encuentran en las "instituciones, prácticas y actitudes sociales", 2) las limitaciones técnicas, y 3) aspectos emergentes en un contexto de uso.

Dado que en el proceso de desarrollo de algoritmos trabajan muchos ingenieros, es difícil (sino imposible) delimitar la responsabilidad que se deriva de la "contaminación" de los algoritmos por el hecho de inconscientemente integrar sus valores en ellos. Dada esta enorme complejidad, no me atrevo a atribuir la responsabilidad a los desarrolladores ¿cómo evitas voluntariamente hacer juicios de valor? En mi opinión los juicios de valor son un aspecto inevitable de la consciencia individual de cada individuo, un evento *ex ante* que, en efecto, sesga nuestras opiniones, preferencias y formas de actuar. Sin embargo, los efectos negativos que pueden llegar a resultar de esto en Facebook (volvamos al caso racial de antes) necesitan de alguien que sea responsable en última instancia, por lo que propongo que, en estos casos, se le atribuya responsabilidad a Facebook como empresa, ya que en última instancia es Facebook quien decide si contratar o no a determinada persona, con todo lo que ello conlleva. Además, Facebook tiene una cultura determinada que se esfuerza por mantener (más del 90% de las donaciones políticas de los empleados de Facebook en la última campaña electoral fueron al Partido Demócrata[218]) y cuadra con la primera fuente de sesgo identificada por Friedman y Nissenbaum (valores sociales preexistentes en las instituciones).

Respecto al sesgo derivado de las limitaciones técnicas, opino que en el caso de Facebook no se suele deber a errores de diseño, sino a errores en los conjuntos de datos procesados. El algoritmo adopta inadvertidamente los defectos de los datos que se han utilizado para entrenarlo, y estos sesgos se ocultan en los resultados producidos (por ejemplo, el sesgo de enseñar más posts de violencia racial a individuos afroamericanos

[218] Facebook, Twitter employees send over 90% of political donations to Democrats: report. URL: https://www.foxnews.com/politics/facebook-twitter-employees-donations-democrats Última consulta 2/2/2022.

por el hecho de que estos suelen tener más *engagement* con contenido que incluye gente de su raza).

Respecto al sesgo emergente, sabemos que se deriva de un cambio en los usuarios o stakeholders de Facebook. En caso de que haya un cambio en un usuario, un algoritmo de aprendizaje automático es capaz de detectar ese cambio (que interpreta como una entrada de datos) y actualizar su sistema de decisiones algorítmica de forma sesgada (negativamente). Pongamos que alguien está deslizándose sobre su feed y accidentalmente (e inadvertidamente) interacciona con un post de violencia racista. El algoritmo interpretará está acción de forma que a ese usuario se le enseñen más a menudo este tipo de posts. En este tipo de caso, me parece que no queda nada claro quién tiene la responsabilidad. No parece que sean los desarrolladores, ya que nada han tenido que ver con la acción (cambio) del usuario, ni Facebook como empresa (aunque aquí entra el debate de si Facebook es un agregador de contenidos que no responde del tipo de contenido que hay en su plataforma o si es una compañía de *Media* que tiene responsabilidad de editar y filtrar contenido). Diría que la responsabilidad es del usuario cuando haga este cambio de forma consciente, y que cuando haga este cambio de forma inconsciente sencillamente estamos ante una situación en la que la responsabilidad queda diluida y no es atribuible a ningún actor en particular.

Análisis aplicado: **resultados injustos**

Respecto a las ideas tratadas en el punto 6.4, cabe decir que la construcción de perfiles a través de algoritmos suele citarse como una fuente de discriminación. Los algoritmos de elaboración de perfiles identifican correlaciones y hacen predicciones sobre comportamientos a nivel de grupo, de tal forma que el usuario es comprendido a partir de las conexiones con otros individuos, y no por su comportamiento real. Con este sistema se puede crear inadvertidamente una base de evidencia que conduzca a la discriminación, y de ahí que merezca especial atención el tratamiento de los atributos sensibles en la elaboración de perfiles a través de algoritmos.

Facebook tiene acceso a datos sobre atributos sensibles de sus usuarios (sexo, raza, edad, tus fotos, con que usuarios has tenido conversaciones, tu número de teléfono y el de tus contactos, tu dirección de IP...). Vemos que es una cantidad ingente de información, y particularmente sensible. Por ejemplo, con lo que Facebook sabe a través

de las fotos de un usuario y la dirección de IP, puede haber un cambio muy acusado en los resultados que produce el algoritmo. De hecho, en 2019, Facebook llevó a cabo un experimento tratando de acallar las críticas de discriminación que estaba recibiendo. Encontraron dos regiones que eran mayoritariamente blancas y dos que eran mayoritariamente negras. Luego probaron para ver cuántos de sus anuncios estaban dirigidos a personas en cada área. Uno de los resultados del experimento es muy llamativo por las asunciones racistas basadas en atributos sensibles que Facebook está haciendo: el 87% de los usuarios que recibieron anuncios sobre Hip Hop eran afroamericanos[219], mientras que solo el 13% eran blancos. En mi opinión, la responsabilidad de esto la tiene Facebook como empresa, ya que está dispuesta a utilizar cualquier fuente de datos por sensible que sea, si utilizarla "mejora" el algoritmo y aumenta los ingresos de la empresa generados a través de anuncios. Los desarrolladores no deberían tener responsabilidad en este caso, ya que realmente no pueden "decidir" si implementar esos atributos sensibles o no (si tienen valor predictivo y deciden unilateralmente no incluir estos atributos en el modelo, Facebook los despide).

También está el problema de que identificar qué atributos son sensibles no es una tarea para nada obvia, especialmente cuando los algoritmos acceden a conjuntos de datos vinculados entre sí. Los perfiles construidos a partir de características neutras, como el código postal, corren el peligro de poder solaparse inadvertidamente[220] con otros perfiles, actuando como indicador para atributos sensibles tales como la raza, el género, la orientación sexual, etc.

Relacionado con esto está la personalización, y los potenciales problemas éticos que de ella se derivan. La personalización puede segmentar una población para que solo algunos segmentos sean merecedores de recibir algunas oportunidades o información, reforzando las (des)ventajas sociales existentes. El ejemplo de los anuncios de Hip Hop expuesto anteriormente sirve para ilustrar este problema ético, aunque sea de una forma "poco trascendente" (no hace daño a nadie recibir anuncios de Hip Hop o dejar de recibirlos). Sin embargo, la personalización también puede tener efectos sobre aspectos más serios de la distribución de oportunidades o información. Ejemplos que se me ocurren

[219] CCN. *Is Facebook racist? Research suggests ad platform uses stereotypes.* URL: https://www.ccn.com/is-facebook-racist-research-suggests-ad-platform-uses-stereotypes/ Última consulta: 2/2/2022

[220] Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.

son la promoción o falta de ella de posts políticos sesgados (sean pro-republicanos o pro-demócratas), de posts de violencia racial o la desinformación, siendo la razón de fondo para que esto sea promovido el hecho de la segmentación de la población en aras de la personalización. La responsabilidad ética en estos casos debería recaer en Facebook como empresa, y no sobre los desarrolladores, ya que es Facebook la que decide utilizar la personalización como una ventaja competitiva o una característica integral de su modelo de negocio.

Análisis aplicado: **Efectos transformadores. Problemas de autonomía y de privacidad informacional**

Respecto a las ideas tratadas en el punto 6.5, referentes a los problemas de autonomía y privacidad informacional, podemos decir que están bastante ligados a lo tratado en previamente en el análisis del 6.4. Los algoritmos de personalización reducen la diversidad de información que encuentran los usuarios, excluyendo contenidos que se consideran irrelevantes o contradictorios con las creencias del usuario. Esto hace que el usuario pierda autonomía, ya que la diversidad de la información es una condición necesaria para que exista la autonomía. Los algoritmos de filtrado crean "cámaras de eco" sin información contradictoria que pueden impedir la autonomía en la toma de decisiones y polarizar políticamente a los individuos de una sociedad. En el caso de Facebook, esto tiene una carga de responsabilidad clara; recordemos el papel que jugó en las elecciones de USA en 2016.

En septiembre de 2017, Facebook reveló que "fuentes" rusas habían pagado $100.000 a cambio de alrededor de 3.000 Facebook Ads de contenido político "candente" (incluyendo derechos LTGBI, derecho a las armas, problemas raciales...) antes de las elecciones de 2016. En su audiencia con el Congreso de USA, Zuckerberg dijo: "La mayoría de los anuncios se compran mediante programación a través de nuestras aplicaciones y sitio web sin que el anunciante hable con nadie en Facebook. Eso es lo que sucedió aquí"[221]. Esos anuncios fueron dirigidos a los usuarios a través de los algoritmos que promocionan cualquier otro contenido en Facebook, en base a la ingente cantidad de datos que Facebook tiene sobre sus usuarios (sean estos identificables o no). Me parece que aquí hay una doble responsabilidad ética por parte de Facebook como empresa, el

[221] Nolan, L. (2018). Medium. *How Facebook influenced the US 2016 Elections.* URL: https://medium.com/nolan-media/how-facebook-affected-the-2016-u-s-election-751125cdf88c

primero referido a la gestión de los datos de los usuarios y como estos interactúan con el algoritmo, y el segundo, la falta de procesos para prevenir que ocurran este tipo de cosas, que como vemos en el ejemplo de las elecciones de 2016, es hasta negligente.

Además, y de forma más preocupante a nivel ético, Facebook como tal ha sido acusada de sesgo interno favorable a políticos demócratas. En una serie de emails filtrados después de las elecciones entre Sheryl Sandberg (BoD de Facebook) y John Podesta (director de campaña para Hillary Clinton), Sandberg ofreció poner a Podesta en contacto con e, Mark Zuckerberg, afirmando que Zuckerberg estaba interesado en influir en las políticas relacionadas con "objetivos orientados a la sociedad (como la inmigración, la educación o la investigación científica básica)"[222]. Podesta parece haber organizado al menos esa reunión; su asistente le envió un correo electrónico en agosto de 2015 con instrucciones para llegar a la oficina de Zuckerberg. Otros correos electrónicos mostraron que se organizó una reunión entre Sandberg y Podesta para "tratar la investigación sobre género y liderazgo de las mujeres"[223]. Hay emails de Sandberg que rezan cosas como: "Espero con interés trabajar con ustedes para elegir a la primera mujer presidenta de los Estados Unidos"[224]. Que una empresa con el poder de influencia algorítmica como es Facebook esté entramada de lleno en la política estadounidense y pueda influir en el resultado de unas elecciones, es un problema ético de calado internacional, y desde mi punto de vista Facebook debería ser responsable como empresa, y los directivos involucrados, responsables de forma personal.

Respecto a la privacidad informacional, opino (como Schermer) que en Facebook quizá no es el modelo más adecuado para atajar los problemas de privacidad derivados del uso de algoritmos, ya que, por cómo hemos visto que funciona el algoritmo de Facebook, una gran parte de la información que te caracteriza la infiere no de tu perfil, sino a través de tus conexiones e interacciones con otra gente y contenido a través de la plataforma. Es decir, saben quién eres hasta un nivel de detalle que supera los límites de

[222] Nolan, L. (2018). Medium. *How Facebook influenced the US 2016 Elections.* URL: https://medium.com/nolan-media/how-facebook-affected-the-2016-u-s-election-751125cdf88c
[223] Nolan, L. (2018). Medium. *How Facebook influenced the US 2016 Elections.* URL: https://medium.com/nolan-media/how-facebook-affected-the-2016-u-s-election-751125cdf88c
[224] Nolan, L. (2018). Medium. *How Facebook influenced the US 2016 Elections.* URL: https://medium.com/nolan-media/how-facebook-affected-the-2016-u-s-election-751125cdf88c

privacidad (puede que el algoritmo te conozca incluso más que gente de tu familia), independientemente de que seas identificable como Pepe Pérez o Sara Martínez.

Van Wel y Royakkers afirman que la construcción de la identidad externa mediante algoritmos es un tipo de des-individualización o una "tendencia a juzgar y tratar a las personas en base a las características del grupo en lugar de en base a sus propias características y méritos individuales". Esto tiene especial calado en Facebook, ya que actualmente estamos viviendo una época política marcada por un concepto conocido como *group politics*, que inevitablemente lleva a la polarización política y es capaz de corroer el substrato social de naciones, e incluso de cosmovisiones globales, como he comentado más arriba.

Análisis aplicado: **Trazabilidad y responsabilidad moral**

Respecto a los desafíos éticos relacionados con la trazabilidad y la responsabilidad moral del punto 6.6, cabe destacar varios en el ámbito de Facebook. El algoritmo de Facebook tiene gran parte de sus algoritmos desarrollados bajo la técnica de aprendizaje automático, en lugar de "manuscrito". Estos algoritmos acarrean desafíos particulares, que desafían la concepción tradicional de la responsabilidad del desarrollador. El modelo tradicional de responsabilidad falla porque "nadie tiene suficiente control sobre las acciones de la máquina para poder asumir la responsabilidad de las mismas"[225]. La brecha entre el control del desarrollador y el comportamiento del algoritmo crea una brecha de responsabilidad en la que la culpa puede asignarse a varios agentes morales simultáneamente, como hemos visto a lo largo del análisis de Facebook.

Siguiendo el razonamiento de varios expertos en el ámbito de "ética de máquinas" (Morek, Floridi y Sanders), puede argumentarse que, en ciertos casos (que se dan en muchos algoritmos de Facebook), los responsables de una acción y sus efectos deberían ser los propios algoritmos, dadas unas características concretas (autonomía significativa, comportamiento interactivo e implicación de responsabilidad causal en sus decisiones). Negar agencia a los agentes artificiales tiene como consecuencia que los desarrolladores acaban siendo responsables del comportamiento poco ético de sus creaciones semiautónomas. El problema de esto es que no parece muy satisfactorio (¿cómo

[225] Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.

responsabilizas de sus acciones a un algoritmo en la práctica?). La responsabilidad queda en tierra de nadie y se diluye detrás de teoría sin aplicación práctica.

Las decisiones éticas requieren que los agentes evalúen la conveniencia de diferentes cursos de acción que presentan conflictos entre los intereses de las partes implicadas. De acuerdo con el testimonio de Megan Haugen, que destapó el escándalo de Facebook en septiembre de 2021 (entre otras cosas, destapando que Facebook tenía datos empíricos sobre la influencia de su producto en niñas adolescentes y un claro empeoramiento de la salud mental) en Facebook, cuando existe este tipo de conflicto entre los intereses de los usuarios y los de Facebook, "Facebook siempre decide en favor de Facebook". Esto da una enorme responsabilidad a Facebook como empresa, y en particular a sus directivos (aquí nada tienen que ver los desarrolladores, ya que el problema ético surge de decisiones estratégicas y no tecnológicas).

Más allá de la naturaleza de la agencia moral en las máquinas, en el campo de ética de las máquinas también se investiga la mejor manera de diseñar razonamiento y comportamientos morales en algoritmos autónomos como agentes morales y éticos. Para delimitar la responsabilidad de Facebook respecto a esto, deberíamos saber si la empresa toma medidas activas para desarrollar comportamientos morales en sus algoritmos autónomos. No he conseguido encontrar nada al respecto, lo que me hace pensar que, en este caso, la responsabilidad se desplaza más del desarrollador a Facebook como empresa, que es negligente y no incluye como parte integral del trabajo de los desarrolladores la atención al desarrollo moral de sus creaciones autónomas.

La automatización de la toma de decisiones crea problemas de coherencia ética entre humanos y algoritmos. El problema es que muchos de los principios éticos utilizados por los responsables humanos son difícilmente computables. Ha habido varias sugerencias para atajar este problema, pero, en última instancia, el desarrollo del carácter de los humanos y el de las máquinas son demasiado distintos para poder compararse. En mi opinión, para Facebook esto significa que no puede evadir la responsabilidad de los efectos derivados de su uso de algoritmos.

8. CONCLUSIÓN

Los algoritmos median cada vez más la vida digital y la toma de decisiones empresariales. El presente trabajo ha tratado de analizar desde la perspectiva de la ética

las implicaciones que tiene el uso de algoritmos en el ámbito empresarial. Para ello, me he servido de: 1) una revisión del debate existente sobre los aspectos éticos de los algoritmos; 2) un mapa para organizar la discusión de las implicaciones éticas desarrollado por Mittelstadt et al., y 3) un caso empresarial para aplicar la teoría desarrollada a lo largo de la investigación al análisis práctico de las implicaciones éticas del uso de algoritmos en Facebook.

La revisión realizada en esta investigación se ha ceñido a la literatura que trata sobre algoritmos, en concreto, a ámbitos de investigación como son los siguientes: algoritmos en el ámbito empresarial, responsabilidad derivada del uso de algoritmos, inteligencia artificial y empresa, diseño ético de algoritmos, empresa y Big Data, responsabilidad del desarrollador en el uso de algoritmos y ética de la Inteligencia Artificial.

El debate sobre un concepto tan abstracto como el de "algoritmo" inevitablemente se encuentra con problemas propios del lenguaje. Dado que por "algoritmo" podemos entender muchos tipos distintos de software[226], es difícil desarrollar una ética general que abarque su uso. A pesar de esta limitación, en la revisión de la literatura han surgido varios hilos conductores que guían el debate ético arrojando luz sobre los principales problemas éticos pertinentes a los algoritmos, y cómo estos afectan en el ámbito de la empresa. .

La articulación de estos hilos conductores en base al mapa ético desarrollado por Mittelstadt et al. ha resultado de gran utilidad para distinguir entre los distintos tipos de problemas éticos generados por los algoritmos. El mapa utilizado demuestra que la resolución de problemas en uno de los 6 aspectos analizados no implica necesariamente una resolución general, que permita hacer el juicio sobre si un algoritmo es efectivamente ético o no. Una decisión algorítmica perfectamente auditable, o que se base en pruebas concluyentes, puede, sin embargo, causar efectos injustos y transformadores, sin que haya formas evidentes de buscar responsables entre la red de personas que contribuyen a esos efectos.[227]

[226] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

[227] Ibídem.

Por todo ello, podemos concluir que es necesaria más investigación (y, potencialmente, regulación *a priori*) en el ámbito de la responsabilidad en el desarrollo y uso de algoritmos a nivel empresarial, ya que, especialmente con empresas que tratan directamente con una masa de usuarios (como Facebook), si aumenta la gravedad o presencia de los efectos negativos derivados del uso de algoritmos, el problema de quién es responsable cobrará todavía más relevancia y necesitará una respuesta por parte de los gobiernos y las propias empresas.

9. DEFINICIÓN DE TÉRMINOS CLAVE

Inteligencia artificial: "La inteligencia artificial (IA) es una disciplina académica relacionada con la teoría de la computación cuyo objetivo es emular algunas de las facultades intelectuales humanas en sistemas artificiales. Con inteligencia humana nos referimos típicamente a procesos de percepción sensorial (visión, audición, etc.) y a sus consiguientes procesos de reconocimiento de patrones, por lo que las aplicaciones más habituales de la IA son el tratamiento de datos y la identificación de sistemas. (...) El diseño de un sistema de inteligencia artificial normalmente requiere la utilización de herramientas de disciplinas muy diferentes como el cálculo numérico, la estadística, la informática, el procesado de señales, el control automático, la robótica o la neurociencia. (...) Un sistema de inteligencia artificial requiere de una secuencia finita de instrucciones que especifique las diferentes acciones que ejecuta la computadora para resolver un determinado problema. Esta secuencia de instrucciones constituye la *estructura algorítmica* del sistema de inteligencia artificial."[228]

Algoritmo: "Se conoce como *método efectivo o algoritmo* al procedimiento para encontrar la solución a un problema mediante la reducción del mismo a un conjunto de reglas."[229]

Procesamiento del lenguaje natural (PNL): "El Procesamiento del Lenguaje Natural es el campo de conocimiento de la Inteligencia Artificial que se ocupa de la investigar la

[228] Benítez, R., Escudero, G., Kanaan, S., & Rodó, D. M. (2014). *Inteligencia artificial avanzada*. Editorial UOC. pp. 10,11.
[229] Benítez, R., Escudero, G., Kanaan, S., & Rodó, D. M. (2014). *Inteligencia artificial avanzada*. Editorial UOC. pp. 11.

manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales, como el español, el inglés o el chino. Virtualmente, cualquier lengua humana puede ser tratada por un ordenador de dos formas: 1) a través de modelos lógicos (desarrollados por lingüistas computacionales en base a patrones estructurales y gramáticos en las lenguas); y 2) a través de modelos probabilísticos basados en datos (los lingüistas computacionales recogen datos de una lengua, y a partir de esos datos se calculan las frecuencias de diferentes unidades lingüísticas y su probabilidad de aparecer en un contexto determinado).[230]

Aprendizaje automático: "El aprendizaje automático (ML) es el proceso mediante el cual se usan modelos matemáticos de datos para ayudar a un equipo a aprender sin instrucciones directas. Se considera un subconjunto de la inteligencia artificial (IA). El aprendizaje automático usa algoritmos para identificar patrones en los datos, y esos patrones luego se usan para crear un modelo de datos que puede hacer predicciones. Con más experiencia y datos, los resultados del aprendizaje automático son más precisos, de forma muy similar a cómo los humanos mejoran con más práctica. La adaptabilidad del aprendizaje automático lo convierte en una excelente opción en escenarios en los que los datos siempre cambian, la naturaleza de la solicitud o la tarea siempre se transforma o la codificación de una solución sería realmente imposible."[232]

Big Data: "Datos de un tamaño muy grande, típicamente en la medida en que su manipulación y gestión presentan importantes desafíos logísticos; (también) la rama de la informática que implica dichos datos."[233]

10. REFERENCIAS

<u>Papers:</u>

[230] Moreno, A. (2018). *Procesamiento del lenguaje natural ¿qué es?* Recuperado el 6 de diciembre de 2021. Instituto de Ingeniería del Conocimiento: https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/

[232] Microsoft Azure (2021). *¿Qué es el aprendizaje automático?* Recuperado el 5 de diciembre de 2021. Microsoft Azure: https://azure.microsoft.com/es-es/overview/what-is-machine-learning-platform/#benefits

[233] Oxford University Press (s.f.). *Big Data*. Recuperado el 5 de diciembre de 2021. Oxford English Dictionary: https://www.oed.com/view/Entry/18833#eid301162177

Ananny M (2016) Toward an ethics of algorithms convening, observation, probability, and timeliness. *Science, Technology & Human Values* 41(1): 93–117.

Barocas, S. (2014). Data mining and the discourse on discrimination. In *Data ethics workshop, conference on knowledge discovery and data mining* (pp. 1-4).

Bello P and Bringsjord S (2012) On how to build a moral machine. *Topoi* 32(2): 251–266.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, *1*, 316-334.

Bozdag E (2013) Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15(3): 209–227.

Brendel, A. B., Mirbabaie, M., Lembcke, T. B., & Hofeditz, L. (2021). Ethical Management of Artificial Intelligence. *Sustainability*, *13*(4), 1974.

Burrell J (2016) How the machine 'thinks:' Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.

Danna A and Gandy OH Jr (2002) All that glitters is not gold: Digging beneath the surface of data mining. Journal of Business Ethics 40(4): 373–386.
Davis (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology MIS Quarterly: Management Information Systems

Davis M, Kumiega A and Van Vliet B. (2013). Ethics, finance, and automation: A preliminary survey of problems in high frequency trading. *Science and Engineering Ethics* 19(3): 851–874.

Ferrell, O. C., & Ferrell, L. (2021). Applying the Hunt Vitell ethics model to artificial intelligence ethics. *Journal of Global Scholars of Marketing Science*, *31*(2), 178-188.

Floridi L (2008) The method of levels of abstraction. *Minds and Machines* 18(3): 303–329.

Floridi L, Fresco N and Primiero G (2014) On malfunctioning software. *Synthese* 192(4): 1199–1220.

Friedman B and Nissenbaum H (1996) Bias in computer systems. *ACM Transactions on Information Systems* (TOIS) 14(3): 330–347.

Granka LA (2010) The politics of search: A decade retrospective. *The Information Society* 26(5): 364–374.

Greene, D., Hoffmann, A. L., & Stark, L. (2019, January). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

Hildebrandt M and Koops B-J (2010) The challenges of ambient law and legal protection in the profiling era. *The Modern Law Review* 73(3): 428–460.

Ioannidis JPA (2005) Why most published research findings are false. *PLoS Medicine* 2(8): e124.

Kim, B., Patel, K., Rostamizadeh, A., & Shah, J. (2015). Scalable and interpretable data representation for high-dimensional, complex data. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1).

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, communication & society*, *20*(1), 14-29.

Kraemer F, van Overveld K and Peterson M (2011) Is there an ethics of algorithms? *Ethics and Information Technology* 13(3): 251–260.

Lazer D, Kennedy R, King G, et al. (2014) The parable of Google flu: Traps in big data analysis. *Science* 343(6176): 1203–1205.

Leese M (2014) The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue* 45(5): 494–511.
Macnish K (2012) Unblinking eyes: The ethics of automating surveillance. Ethics and Information Technology 14(2): 151–167.

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of business ethics*, *160*(4), 835-850.

Matthias A (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183.

Miller, B. y Record, I. (2013). Justified belief in a digital age: On the epistemic implications of secret Internet technologies. *Episteme* 10(2): 117–134.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

Müller, V. C. (2021). Ethics of artificial intelligence 1. In *The Routledge social science handbook of AI* (pp. 122-137). Routledge.

Ormond, E. (2019). The Ghost in the Machine: The Ethical Risks of AI. *Ormond, E.(2020)'The Ghost in the Machine: The Ethical Risks of AI', The Thinker*, *83*(1), 4-11.

Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry.*

Schermer BW (2011) The limits of privacy in automated profiling and data mining. *Computer Law & Security Review* 27(1): 45–52.

Simon J (2015) Distributed epistemic responsibility in a hyperconnected era. In: Floridi L (ed.) *The Onlife Manifesto*. Springer International Publishing, pp. 145–159.

Tonkens R (2012) Out of character: On the creation of virtuous machines. *Ethics and Information Technology* 14(2): 137–149.

Turilli M and Floridi L (2009) The ethics of information transparency. *Ethics and Information Technology* 11(2): 105–112.

Valiant LG (1984) A theory of the learnable. Communications of the Journal of the ACM 27: 1134–1142.

Van Wel L and Royakkers L. (2004). Ethical issues in web data mining. *Ethics and Information Technology* 6(2): 129–140.

Vellido, A., Martín-Guerrero, J. D., & Lisboa, P. J. (2012). Making machine learning models interpretable. In *ESANN* (Vol. 12, pp. 163-172).

Wiltshire TJ (2015) A prospective framework for the design of ideal artificial moral agents: Insights from the science of heroism in humans. *Minds and Machines* 25(1): 57–71.

Wright, S. A., & Schultz, A. E. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons*, *61*(6), 823-832.

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118-132.

Podcasts:

Ethics of AI Lab, University of Toronto (anfitrión). (2019-presente). Michael Kearns, the ethical algorithm. (Podcast). Spotify: https://open.spotify.com/episode/5EdJY1vYPlelNriH3ThCdS?si=-ppwjgKTQiqKbl5P7fAAAg

Fridman, L. (anfitrión). (2018-presente). *Michael Kearns: Algorithmic Fairness, Bias, Privacy and Ethics in Machine Learning.* (Podcast). Spotify: https://open.spotify.com/episode/3QS8iWIQxczZ4dsg8SKwqh?si=Bo6At4L4QnqBhjD5vl0VaQ

Books:

Bostrom (2014) Superintelligence: Paths, Dangers, Strategies (Book)

Brian Christian (2020) The alignment problem.

Brynjolfsson and McAfee (2014) The Second Machine Age: Work, Progress, and Prosperity in A Time of Brilliant Technologies (Book)

Goldberg (1989) Genetic Algorithms in Search, Optimization and Machine Learning (Book)

Illari PM and Russo F (2014) Causality: Philosophical Theory Meets Scientific Practice. Oxford: Oxford University Press.

James G, Witten D, Hastie T, et al. (2013) An Introduction to Statistical Learning. Vol. 6, New York: Springer.

Kurzweil (2005) The Singularity Is Near: When Humans Transcend Biology (Book)
Pasquale, F. (2015). *The black box society*. Harvard University Press.

Rogers (1995) Diffusion of Innovations (Book)

Russell and Norvig (1995) and subsequent edition Russell and Norvig (2016) Artificial Intelligence: A Modern Approach (Book).

Turing (1950) Computing machinery and intelligence

Zadeh (1965) Fuzzy sets Information and Control

Webs:

Adler P, Falk C, Friedler SA, et al. (2016) Auditing black-box models by obscuring features. arXiv:1602.07043 [cs, stat]. URL: http://arxiv.org/abs/1602.07043

Allen C, Wallach W and Smit I (2006) Why machine ethics? Intelligent Systems, IEEE 21(4) URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1667947 Recuperado el 30 de diciembre de 2021

Barocas S and Selbst AD (2015) Big data's disparate impact. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. URL: http://papers.ssrn.com/abstract=2477899 Recuperado 28 de diciembre de 2021.

Datta A, Sen S and Zick Y. (2016). Algorithmic transparency via quantitative input influence. In: Proceedings of 37th IEEE symposium on security and privacy, San Jose, USA. URL: http://www.ieee-security.org/TC/SP2016/papers/0824a598.pdf

McKinsey. (2021) *The State of AI in 2021*: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2021

Orseau, L. y Armstrong, S. (2016). *Safely interruptible agents*. [Archivo PDF]. URL: http://intelligence.org/files/Interruptibility.

Tene O and Polonetsky J (2013a) Big data for all: Privacy and user control in the age of analytics. URL: http://heinonlinebackup.com/hol-cgibin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20 Recuperado el 28 de diciembre.

Zarsky T (2013) Transparent predictions. University of Illinois Law Review 2013(4). URL: http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2324240 Recuperado 28 de diciembre de 2021.