IUS ET SCIENTIA
2024 • Vol. 10 • Nº 2 • ISSN 2444-8478
https://editorial.us.es/es/revistas/ius-et-scientia
https://dx.doi.org/10.12795/IETSCIENTIA • © Editorial Universidad de Sevilla 2024

La construcción de un algoritmo «ético»

THE CONSTRUCTION OF AN 'ETHICAL' ALGORITHM

Alessandra Esther Castagnedi Ramirez
Universidad de Sevilla
Alessandrae.castagnediramirez@gmail.com 0000-0002-4905-0362

Recibido: 11 de septiembre de 2024 | Aceptado: 08 de diciembre de 2024

RESUMEN

El artículo aborda la ética en la inteligencia artificial (IA), destacando el riesgo de sesgos discriminatorios en los algoritmos que afectan decisiones cruciales, como la aprobación de hipotecas o la asignación de atención médica. A partir de 2016, se incrementó la participación de gobiernos y organizaciones en el debate sobre la creación de IA ética y justa. Ejemplos de sesgos algorítmicos ilustran problemas de racismo y exclusión, donde algoritmos, incluso sin información explícita de raza o género, perpetúan inequidades sociales. Además, se analiza la necesidad de transparencia y "explicabilidad en los sistemas de IA para asegurar decisiones comprensibles y auditables. Propone un marco legal europeo que garantice la equidad y minimice los sesgos, sugiriendo certificaciones y una estricta regulación en sistemas de alto riesgo, señalando que la ética debe ser integrada en el diseño y aplicación de estas tecnologías para evitar abusos y discriminaciones.

PALABRAS CLAVE

Ética en Inteligencia Artificial (IA)
Sesgos Algorítmicos
Transparencia
Explicabilidad
Regulación Europea
Toma de Decisiones
Automatizadas
Derechos Humanos

ABSTRACT

The article addresses ethics in artificial intelligence (AI), highlighting the risk of discriminatory bias in algorithms that affect crucial decisions, such as mortgage approval or health care allocation. Since 2016, governments and organisations have become increasingly involved in the debate on creating ethical and fair AI. Examples of algorithmic biases illustrate problems of racism and exclusion, where algorithms, even without explicit race or gender information, perpetuate social inequities. Furthermore, it analyses the need for transparency and explainability in AI systems to ensure understandable and auditable decisions. It proposes a European legal framework that guarantees fairness and minimises bias, suggesting certifications and strict regulation in high-risk systems, noting that ethics must be integrated into the design and application of these technologies to avoid abuse and discrimination.

KEYWORDS

Ethics in Artificial Intelligence (AI)
Algorithmic Bias
Transparency
Explainability
European Regulation
Automated Decision-Making
Human Rights

I. Introducción: Reflexiones éticas en torno a los algoritmos

Nick Bostrom, en un destacado artículo publicado en la prestigiosa revista «The Cambridge Handbook of Artificial Intelligence», dedica una sección del elaborado a la importancia de la ética en el contexto de la IA. La relevancia de este aspecto nace desde las reflexiones más profundas que han surgido cuando algunos problemas, como la reproducción de sesgos discriminatorios (Degli Esposti, 2023, p. 33), se convirtieron en escándalos que empezaron a preocupar la opinión pública.

La investigación en este campo ha conocido un desarrollo importante a partir del 2016, año en el que los gobiernos nacionales, las organizaciones no gubernamentales y las empresas privadas empezaron a cubrir un rol relevante en el gran debate sobre la IA y los algoritmos «justos» y «éticos» (Ochigame, 2019)[1]. Los algoritmos junto a los datos juegan un papel relevante, activo y participativo, en la implementación y en la aplicación de las consecuencias dañinas de las nuevas tecnologías. El ejemplo subyacente muestra con claridad una de las maneras con las que los algoritmos puedan afectar la realidad a través de su uso en el ámbito de las concesiones de hipotecas:

> Imagine, in the near future, a bank using a machine learning algorithm to recommend mortgage applications for approval. A rejected applicant brings a lawsuit against the bank, alleging that algorithm is discriminating racially against mortgage applicants. The bank replies that this is impossible since the algorithm is deliberately blinded to the race of the applicants. Indeed, that was part of the bank's rationale for implementing the system. Even so, statistics show that the bank's approval rate for black applicants has been steadily dropping. Submitting ten apparently equally qualified genuine applicants (as determined by a separate panel of human judges) shows that the algorithm accepts white applicants and rejects black applicants. What could possibly be happening? (Bostrom, Yudkowsky, 2014, p. 1)[2].

1. La concienciación sobre la importancia de implementar la investigación en el nuevo campo emergente de los algoritmos «éticos» tiene base de partida distintas, si se analiza el trayecto que ha sido cumplido por la Unión europea o por los EE. UU. Sin embargo, el anillo que asumió el rol de enlace entre estas dos distintas realidades se concretizó con el reconocimiento de Joichito como «experto en ética de la IA» por parte del presidente Barack Obama, diseminando a nivel global el discurso sobre la ética de la IA e integrando también todas aquellas políticas que a nivel europeo ya se estaban desarrollando, poniendo al centro de su investigación la importancia de los derechos humanos fundamentales. En este aspecto, relevante el artículo escrito por parte de Rodrigo Ochigame, investigador de prestigiosas instituciones con el MIT Media Lab, que permite al lector tener un cuadro panorámico respecto a la evolución del papel de la ética a partir de 2016 en el panorama estadunidense y la respuesta que fue dada por parte de los subjetos que interactúan en la sociedad frente a una posible manipulación realizada por media de las nuevas tecnologías.

2. El texto traducido aparece de la siguiente forma: «Imaginemos que, en un futuro próximo, un banco utiliza un algoritmo de aprendizaje automático para recomendar la aprobación de solicitudes de hipotecas. Un solicitante rechazado interpone una demanda contra el banco, alegando que el algoritmo discrimina racialmente a los solicitantes de hipotecas. El banco responde que eso es imposible, ya que el algoritmo ignora deliberadamente la raza de los solicitantes. De hecho, esa fue parte de la justificación del banco para implantar el sistema. Aun así, las estadísticas muestran que la tasa de aprobación del banco para los solicitantes negros ha ido disminuyendo constantemente. La presentación de diez solicitantes auténticos aparentemente igual de cualificados (según lo determinado por otro panel de jueces humanos) muestra que el algoritmo acepta a los solicitantes blancos y rechaza a los negros. ¿Qué puede estar pasando?».

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

II. Algunos desafíos éticos en la salud

Este fenómeno no es el único que demuestra los evidentes aspectos críticos que derivan del uso de los algoritmos. La misma situación se ha repetido en diversos contextos, entre ellos también en el ámbito hospitalario. Las estructuras sanitarias de los EE. UU., por ejemplo, se sirven de un sistema de IA, cuya función es determinar, atreves de operaciones de predicción, cuales pacientes merecen ser destinatarios de una determinada cura (más apropiada y de valor económico superior) y cuáles no lo son. Estos algoritmos predictivos son una evidente prueba de la importancia que recubre la ética en el uso de la IA y de las reflexiones en torno al respeto de los derechos humanos fundamentales como, en este caso específico, el derecho a la no discriminación, ya que la mayoría de las veces han reconocido el derecho a recibir asistencia médica sólo a pacientes de piel blanca, descartando los pacientes de piel oscura que igualmente necesitaban la misma intervención médica. Es este el resultado de un proyecto de investigación, realizado entre el año 2013 y 2015, por parte de cuatro profesionistas académicos pertenecientes a Departamentos de *Public Health* de distintas universidades americanas (Obermayer, Powers et al., 2019, pp. 447-453) [3].

Otra víctima de la errata predicción de los algoritmos, a pesar de su grave condición de salud, es Judith Sullivan, mujer de 76 años, que fue dimitida forzosamente del hospital tras la decisión de una de las empresas de seguros médicos más importante del Connecticut, denominada *United Healthcare*. La señora podía caminar pocos pasos, incluso con asistencia, y no podía subir las escaleras hasta la puerta de su casa. Su estado de salud se agravo por la presencia de una herida quirúrgica, cuyo tratamiento preveía cambios de vendaje diarios, a los cuales tras la decisión algorítmica ya no pudo más acceder[4]. Esta elección fue definida por la misma empresa, la cual, sin realizar ningún tipo de investigación concreta sobre sus condiciones patológicas, decidió confiar directamente en la predicción del algoritmo Nh Predict. La función de este último consta en la determinación del diagnóstico de los pacientes, atreves de un estudio del relativo registro médico y de otros factores cuales la atad y las condiciones de salud preexistentes. Esta predicción permite establecer con antelación la cura individual necesaria, el

3. Interesante el resultado mostrado en la Fig. 2 de la publicación científica, realizada por los mismos académicos. El análisis toma en consideración cinco distintas patologías: gravedad de la diabetes, presión arterial alta, insuficiencia renal, colesterol y anemia. Todos estos casos presentan el mismo sesgo: los afroamericanos son menos saludables respecto a los blancos y la diferencia, expresada en términos de porcentaje, es muy alta.

4. Una atenta lectura del informe de Judith Sullivan nos permite estudiar las limitaciones y los desafíos frente a la predicción realizada a través del algoritmo «Nh Predict». Si, por un lado, el algoritmo puede procesar grandes cantidades de datos rápidamente, proporcionando recomendaciones, reduciendo la variabilidad y mejorando la calidad general del cuidado; por el otro, este caso específico muestra como el algoritmo no considera la necesidad del cuidado de las heridas y otros factores como la incapacidad en subir las escaleras. El otro límite evidente es que el mismo informe no proporciona soluciones específicas frente al posible riesgo de hospitalización, correspondiente al 52,5%. A estas faltas, se suma el problema de la ilegibilidad del proceso de toma de decisión, generando de esta forma desconfianza y dificultad de identificación y corrección de errores o sesgos en el sistema. Este hecho real ofrece una visión panorámica de los grandes riesgos que corre la humanidad frente a la «demasiada» confianza que se le atribuye a los algoritmos, dejando en mano de las tecnologías el poder de decisión sobre la vida de los seres humanos y con ello la violación de múltiples derechos humanos.

IUS ET SCIENTIA • 2024
Vol. 10 • Nº 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

tiempo de curación previsto y por ende la fecha de autorización del alta hospitalaria. El problema aquí descrito se empezó a tomar en serio tras la muerte[5] de dos pacientes a los cuales se les había negado definitivamente la cobertura del seguro. En esta y en otra ocasión se demostró que el 90% de los casos la decisión predictiva resultaba contrarias a las correspondientes dictadas por un ser humano.

La difusión de la información sobre los efectos negativos tecnológicos requiere que cada uno de nosotros amplíe el conocimiento y la comprensión de las consecuencias que derivan del uso de la AI.

Realizando una comparación entre los algoritmos y los vehículos que utilizamos en nuestro día a día, se podría perfectamente averiguar que operan de manera análoga: ambos para poder cumplir su propósito requieren un determinado insumo. En el caso de los coches, este último está representado por el petróleo o el diésel; para los algoritmos, los elementos que posibilitan la obtención de ciertos resultados, ya sea en la toma de decisiones como en las predicciones, son los datos de entrenamiento.

El objetivo programado en cada *software* de IA depende de un conjunto de operaciones y de cálculos matemáticos que los algoritmos realizan a través de unos datos iniciales (*input*) para el desarrollo de otros datos finales (*output*). Utilizando una definición más específica, así como determinado por Robin Hill, el algoritmo es «una estructura de control finita, abstracta, eficaz, compuesta, dada imperativamente, que realiza un propósito determinado en determinadas condiciones» (Hill, 2016, p. 47) En otras palabras, se puede definir «algoritmo» como «el procedimiento de cálculo que consiste en realizar una serie ordenada y finita de instrucciones con algunos datos específicos para encontrar una solución al problema planteado». Dichos algoritmos deben ser anónimos en su proceso de elaboración, a efectos de proteger a la ciudadanía; aunque, por el otro lado, la «opacidad» de los mismos conlleva problemas más amplios. Se patria decidir, en cierta forma, que tienen «vida propia» y esto suele ser la «excusa perfecta» cuando incurren en discriminación y otros errores (Cerrillo i Martínez, 2019).

Por esta razón, no podemos desconocer los retos a los que se enfrenta el derecho para responder a los problemas generados por la IA. Seguramente los algoritmos y la información son la base esencial del progreso en la sociedad contemporánea y los, ciber-ciudadanos son conciencies de ello ya que interactúan, por ejemplo, directamente con sistemas de recomendación algorítmica para elegir una canción, una película, un producto, la afinidad con otro individuo, etc. Del mismo modo, también instituciones como hospitales, colegios, organismos públicos locales, gobiernos nacionales usan siempre más a menudo las predicciones algorítmicas para la toma de decisiones. No obstante, aunque su aplicación ofrezca múltiples ventajas, es crucial es de relavan tía crucial considerar la imparcialidad ética. Este último aspecto pasa desapercibido frente a los ojos de los ciudadanos, que de lo contrario muestran una confianza ciega en los algoritmos, los cuales hoy en día filtran y organizan el mundo que percibimos y la realidad en la que vivimos,

5. Esta información ha sido reportada en dos periódicos digitales, respectivamente *KFF Health News* y *Corriere della Sera,* en los siguientes enlaces: https://www.corriere.it/tecnologia/cards/come-l-intelligenza-artificiale-potrebbe-sfuggire-di-mano-e-causare-danni-dai-sistemi-di-sorveglianza-all-uso-in-guerra/gli-errori-nelle-assicurazioni-sanitarie.shtml?refresh_ce; https://kffhealthnews.org/news/article/biden-administration-software-algorithms-medicare-advantage/.

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

realizando una labor descriptiva de todo lo que nos rodea, tanto desde un punto de vista subjetivo como, los objetivo. Siguiendo su propia lógica, se deduce, por lo tanto, que, en la mayoría de las ocasiones, las ideas de la comunidad humana están guiadas por los algoritmos predictivos y descriptivos de la de la relatad, así como sus acciones y comportamientos. Esta situación se ha fomentado por la capacidad de resolución de los problemas por parte de los algoritmos, cuya lógica esta guida por el individualismo egoísta natural de los lobbies que tienen el poder de definir lo que es considerable verdadero o justo para la humanidad (García - Marzà, 2023, pp. 99-114).

La convicción de que el mundo representado por los algoritmos, tanto a nivel epistemológico como moral, sea neutral y objetivo, así como destacan Astrid Mager y Tarleton Gillespie, es fruto de la mirada ingenua y despistada del ciudadano medio (Mager, 2012, pp. 769-787; Gillespie, 2016, pp. 64-87)[6]. Esta percepción contrasta con la preocupación de los especialistas en la materia, quienes analizan de manera crítica los beneficios y riesgos asociados. La razón se debe principalmente a las facultades que se están reconociendo a los algoritmos, de forma progresiva y en acorde con el paso del tiempo, cuya capacidad supera las competencias humanas en el cálculo, en la predicción y en la toma de decisiones eficientes y eficaces, alcanzando así una sustitución integral de nuestra capacidad de decidir y actuar.

La atención que los *softwares* de IA han captado se debe precisamente al poder que ejercen sobre nuestra toma de decisiones, ya que seleccionan la información que consideran más relevante para cada uno de nosotros, según un atento estudio individual. Tal como se mencionó en un discurso, pronunciado durante la conferencia Governing Algorithms, celebrada el 16 y 17 de mayo de 2013 en Nueva York, «algorithms are invoked as powerful entities that govern, judge, sort, regulate, classify, influence, or otherwise discipline the world» (Barocas, Hood, et alii, 2013,1-12).

Surge espontanea la pregunta siguiente: ¿cómo podemos justificar y auditar las decisiones tomadas por algoritmos, especialmente cuando se basan en técnicas complejas como redes neuronales y aprendizaje automático no supervisado?

La cuestión que deberíamos plantearnos, por lo tanto, es: ¿en qué momento llegamos a pensar y creer firmemente que «razonar para convencer» es lo mismo que «calcular», a la luz de que «cualquier algoritmo puede requerir escrutinio»?

La respuesta está inscrita en nuestra conciencia que nos orienta hacia lo que está bien, poniéndonos en guardia respecto a todo lo que es paralelamente opuesto. Distintas son las dudas que surgen en estas ocasiones, entre ellas si es éticamente aceptable o no crear maquinas inteligentes; si la finalidad planteada sea realmente el progreso para un futuro mejor; si se puede aceptar la introducción de *software* en nuestras vidas, a pesar de la existencia de eventuales riesgos, como pasa en cada investigación que no se realice bajo el pleno control de una supervisión humana, incluso cuando los mencionados riesgos sean menores que los beneficios (Pinto Fontanillo, 2020, 40-41).

6. Se trata de investigadores, cuya actividad se centra en el análisis de la intersección entre tecnología, sociedad, política, e impactos sociales de las plataformas digitales, que en algunos de sus trabajos consideraban que los algoritmos «más que meras herramientas, son también estabilizadores de confianza, garantías prácticas y simbólicas de que sus evaluaciones son justas y precisas, libres de subjetividad, error o intentos de influencia».

IUS ET SCIENTIA • 2024
Vol. 10 • Nº 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

Podríamos seguir hasta el infinito creando un *cocktail* de dudas y preguntas para la difusión e implementación de la carrera de desarrollo de la IA. La experiencia siempre nos ha mostrado que el avance de las nuevas tecnologías va más deprisa que los requerimientos éticos vigentes en la actualidad. De aquí, es entendible la razón por la cual diversos informes y estudios expresan inquietudes sobre las consecuencias del uso de estos sistemas complejos y destacan una preocupación más amplia, relacionada con los potenciales poderes y las posibles formas de abuso o manipulación que estas mismas prácticas conllevan.

Los diversos enfoques y preocupaciones de los analistas se centran, por lo tanto, en la temática de la opacidad algorítmica. La necesidad de garantizar la transparencia se menciona frecuentemente como un elemento crucial para manejar este nuevo orden, ya que la imposibilidad de «acceder a características críticas de los procesos de toma de decisiones» de los algoritmos nos aleja de la construcción de una sociedad basada en un modelo de Estado de Derecho, cuya prioridad es la protección de los derechos humanos fundamentales de los ciudadanos.

Es muy complicado el alcance de la perfección, sobre todo si nos planteamos una visión negativa del uso de la IA. Por eso, así como reportado por el profesor Llano Alonso en su última obra, la solución que debería de avanzar es la instauración de un sistema constitucionalista hibrido que incorpore los valores, principios y derechos constitucionales en la fase de diseño del lenguaje de las maquinas. En otras palabras, este resultado solo se obtiene reconociendo por encima de todo la protección del sistema educativo, formando los futuros técnicos respecto a la importancia de los principios generales, de la protección de los datos personales, del rol de la dignidad humana y de la explicabilidad de los algoritmos en la toma de decisiones (Llano Alonso, 2024, 207).

En este sentido, resulta interesante la propuesta de Andrés Boix Palop, quien, reconociendo la superación del paradigma deductivo racionalista mediante el recurso al cálculo probabilístico y las inferencias algorítmicas (es decir, con otras palabras, el uso de algoritmos en lugar de decisiones humanas), asume que las reglas tradicionales del juego también deben cambiar, así como el papel que recubre el Derecho y, más en específico, la rama del Derecho público. Esta concienciación surge de la integración de la IA en la toma de decisiones típicamente humanas respecto a la actuación administrativa del futuro. Por ello, el autor argumenta que los algoritmos empleados por las administraciones públicas deben ser considerados como reglamentos debido a la función normativa a ellos otorgada. A partir de esta premisa, el autor propone que se apliquen las mismas garantías jurídicas que se utilizan en las normativas tradicionales, tales como la participación ciudadana, la publicidad del código fuente, la evaluación *ex ante* detallada sobre el impacto potencial del algoritmo en los derechos y garantías ciudadanas, y la evaluación *ex post* que asegure de manera continua que el sistema ha operado de forma justa y eficiente, eliminando cualquier error o sesgo que haya podido ser generado. Asimismo, se garantizaría públicamente la accesibilidad del código fuente en todo momento (Boix Palop, 2020, pp. 223-270).

En este aspecto, predictibilidad y transparencia van de la mano en la cultura algorítmica, desempeñando ambos un papel fundamental. Nick Bostrom y Eliezer Yudkowsky destacan esta relación en una de sus publicaciones científicas, señalando que:

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

> La transparencia no es la única característica deseable de la IA. También es importante que los algoritmos sean predecibles para las personas a quienes afectan. La predictibilidad asegura que las decisiones tomadas por IA sean comprensibles y esperables, lo cual es vital para que las personas puedan confiar en estos sistemas. La analogía con el principio legal de *stare decisis* se utiliza para resaltar cómo la predictibilidad en la toma de decisiones legales permite a los ciudadanos optimizar sus vidas dentro de un marco estable. De manera similar, los algoritmos de IA deben ser predecibles para que los sujetos puedan entender y anticipar sus decisiones y comportamientos (Bostrom & Yudkowsky, 2011, p. 2).

Esta perspectiva destaca la estricta relación entre predicción y explicabilidad. La explicabilidad, a su vez, puede desempeñar un rol crucial para mejorar el aprendizaje automático y para limitar los problemas en los sistemas de IA y en los datos de entrenamiento.

Sin embargo, como ha señalado el *National Institute of Standards and Technology* (NIST), un organismo de normalización técnica, «la transparencia no garantiza la explicabilidad, especialmente si el usuario no comprende los principios técnicos». El NIST además subraya riesgos adicionales asociados con la explicabilidad, como la falta de coherencia en las explicaciones generadas por los sistemas y la imposibilidad de corrección humana. Una regla general es que «cuanto más opaco es un modelo, menos explicable se considera».

La Recomendación sobre la Ética de la IA de 2021, publicada por la UNESCO, considera que la explicabilidad es esencialmente «la inteligibilidad de la entrada, salida y funcionamiento de cada componente algorítmico y la forma en que contribuye a los resultados de los sistemas» (UNESCO, 2021, no. 40)[7].

Según el Grupo de Altos Expertos de la Comisión Europea (HLEG) (no. 53), «la explicabilidad se refiere a la capacidad de explicar tanto los procesos técnicos de un sistema de IA como las decisiones humanas relacionadas.» Además, señalan que "la explicabilidad técnica exige que las decisiones tomadas por un sistema de IA sean comprensibles para las personas y que estas puedan rastrearlas. (HLEG, 2018,18).

El HLEG considera que, en algunas ocasiones, la precisión del sistema puede requerir un sacrificio en la explicabilidad, «además, puede que sea necesario buscar un equilibrio entre la mejora de la explicabilidad de un sistema (que puede reducir su precisión) o una mayor precisión de este (a costa de la explicabilidad).»

El enfoque inicial es que «el grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado erróneo o inadecuado» (no. 53). Siguiendo esta línea, se detalla la importancia de la explicabilidad en función del «impacto significativo en la vida de las personas.» De este

7. Es relevante citar la afirmación de la Recomendación UNESCO 2021, número 40: «La explicabilidad implica hacer comprensibles los resultados de los sistemas de IA y proporcionar información al respecto. Esto abarca la claridad sobre la entrada, salida y funcionamiento de cada componente algorítmico y su contribución a los resultados finales. Por lo tanto, la explicabilidad está vinculada con la transparencia, ya que tanto los resultados como los procesos que los generan deben ser comprensibles y rastreables, según el contexto. Los responsables de la IA deben comprometerse a asegurar que los algoritmos desarrollados sean explicables. En aplicaciones de IA donde el impacto en el usuario final no es temporal, es difícilmente reversible o implica bajo riesgo, se debe garantizar una explicación clara de cada decisión que haya llevado a la acción tomada, para que el resultado se considere transparente».

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

modo, la explicabilidad se define como «una explicación adecuada del proceso de toma de decisiones del sistema de IA» que debe "adaptarse al nivel de especialización de la parte interesada.

El profesor Gutiérrez David establece una distinción entre la transparencia y la explicabilidad en los sistemas de IA. Según su análisis, la transparencia intrínseca de un modelo es una característica pasiva, que permite al observador humano comprender el sistema, mientras que la explicabilidad es una característica activa, que implica la generación de explicaciones sobre el comportamiento del modelo (Gutiérrez, 2021, p. 55).

Estas explicaciones incluyen información sobre los datos utilizados, los resultados obtenidos y el proceso completo de toma de decisiones. Por tanto, si la transparencia es el objetivo final, las explicaciones se convierten en las herramientas necesarias para lograr la interoperabilidad del modelo. Es evidente que las explicaciones no deben ser iguales para todos, ya que varían dependiendo de si el receptor es un experto individual o un grupo colectivo.

Para garantizar que la función explicativa se lleve a cabo correctamente, se utiliza un criterio intermedio que permita que las explicaciones sean:

1. Comprensibles para la mayoría de los usuarios,
2. Rigurosas y precisas para garantizar su fiabilidad (Ortiz de Zárate Alcarazo, 2022, p. 338).

Estos conceptos son examinados directamente por el Grupo de Altos Expertos de la Comisión Europea, que propusieron una lista de verificación para la transparencia, incluyendo aspectos como la trazabilidad, la comunicación y la explicabilidad. Para garantizar esta última, se plantea evaluar si las decisiones y los resultados del sistema de IA son «comprensibles» y si se pueden explicar «las razones por las que un sistema adoptó una decisión determinada.» Además, debe ser posible proporcionar a los usuarios una explicación de los resultados específicos. También se cuestiona si el sistema fue diseñado «desde el principio para la interpretabilidad». El HLEG afirma que la transparencia «guarda una relación estrecha con el principio de explicabilidad e incluye la transparencia de los elementos pertinentes para un sistema de IA: los datos, el sistema y los modelos de negocio» (HLEG, 2018, 22). El concepto de transparencia tiene dos significados diferentes no compatibles entre ellos y potencialmente confundibles por parte de la comunidad. Si nos referimos a la ética de la información y, más en general, a la disciplina de gestión de esta como en el presente caso, la «transparencia» suele referirse a formas de hacer visible la información, la cual se logra al reducir o eliminar obstáculos. Específicamente, la transparencia se refiere a la posibilidad de acceder a las informaciones, intenciones o comportamientos que han sido revelados deliberadamente a través de un proceso de divulgación. De lo contrario, en el ámbito puramente informático, el concepto de «transparencia» corresponde a la condición de la cual la información es invisible para el usuario.

En este contexto, la acepción que se utiliza es la primera, que coincide con el significado relacionado a la accesibilidad del tipo de información que un proveedor debe reconocer hacia ciertos agentes. De hecho, el lado opuesto de la medalla, ósea la dificultad de acceso a la información, conduce a una falta di «fiabilidad» aunque sea justificada

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

por: la presencia di vínculos jurídicos en términos de propiedad intelectual; los límites de la misma tecnología respecto al diseño o al ocultamiento de los datos subyacentes; la imposibilidad cognitiva de los humanos en la interpretación de gigantescos modelos algorítmicos y conjuntos de datos; la falta de herramientas adecuadas para visualizar y hacer un seguimiento de grandes volúmenes de código y datos y, finalmente, por el código y los datos mal estructurados imposibles de leer.

Ahora bien, si por un lado la transparencia es uno de los elementos fundamentales para la instalación de una cultura algorítmica ética, deberíamos de reflexionar si realmente la misma es clasificable en el mundo de los algoritmos como principio o si resulta, así como nos sugiere Luciano Floridi, una condición pre-ética para consentir o frenar otras prácticas y principios éticos.

Si consideramos la transparencia como un valor y principio fundamental en instituciones, productos y servicios a la par de otros valores constitucionales, cuales la libertad, la igualdad, la justicia y el pluralismo político, se exigiría que el contenido de la Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno, cuyo objetivo es la integración de la legislación ya existente en materia, contenida en el artículo 105.b) del texto constitucional español; en el artículo 37 la Ley 30/1992, de 26 de noviembre, de Régimen Jurídico de las Administraciones Públicas y del Procedimiento Administrativo Común; en la Ley 27/2006, de 18 de julio, por la que se regulan los derechos de acceso a la información, de participación pública y de acceso a la justicia en materia de medio ambiente y en la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, que regula el uso privado de documentos en poder de Administraciones y organismos del sector público, también se aplicara en un contexto algorítmico, sin tomar en consideración las eventuales consecuencias negativas.

A veces la opacidad puede ser más útil, por ejemplo, para asegurar la confidencialidad de las preferencias y votos políticos de los ciudadanos o para garantizar la competencia en las subastas de servicios públicos. De hecho, incluso en contextos algorítmicos, la completa transparencia puede causar problemas éticos específicos. Puede proporcionar a los usuarios información relevante sobre las características y limitaciones de un algoritmo, pero también puede sobrecargar a los usuarios con información y, de ese modo matizar un algoritmo. Por ejemplo, el debate sobre la prioridad de la transparencia es particularmente controvertido en el contexto sanitario, en el cual los algoritmos facilitan el acceso a registros médicos electrónicos que, si por un lado ayudan a la investigación en implementar técnicas y herramientas salvavidas, por el otro lado simultáneamente puede exponer a los pacientes al fraude o a una violación de la privacidad, ya que se divulga su información personal. Por esta razón se necesita investigar sobre la naturaleza de la transparencia y las implicaciones éticas que la misma tiene una vez concedida. De hecho, la transparencia también puede permitir a las personas la posibilidad de engañar al sistema. El conocimiento de la fuente de un conjunto de datos, de los asuntos por los cuales se realizó el muestreo o de las métricas utilizadas por un algoritmo para clasificar nuevas entradas, puede ser usado para entender cómo explotar un algoritmo (Floridi, 2019, p. 155).

La clave por lo tanto es el análisis, no tanto de los canales que permiten el acceso a la información, sino la evaluación de sus características éticas del proceso de producción

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

de esta en el contexto de organizaciones heterogéneas, donde los agentes humanos y artificiales son partes de un solo sistema general. La clarificación de este aspecto conduce a un cambio en la forma en que debe respaldarse la transparencia. La acción de divulgación de solo códigos éticos o profesionales, grabaciones o resúmenes de actividades, actas o informes de reuniones disminuye en comparación con los efectos de hacer públicamente visibles también los detalles de producción, elaboración e interpretación de dicha información. Este cambio no es sorprendente una vez que se aprecia que la transparencia no es un principio ético en sí mismo.

La transparencia de la información, como se definió anteriormente, es una condición pro-ética que se convierte en una herramienta valiosa para descubrir los principios que idealmente inspiran las decisiones de las organizaciones y aquellos principios que se respaldan fácticamente en sus actividades cotidianas. De esta manera se entiende perfectamente la distinción entre los cincos principios fundamentales de la IA (beneficencia, no maleficencia, autonomía, justicia y explicabilidad) y la transparencia, que funciona de indicador ético respecto a las practicas que se están poniendo en marcha.

El panorama se vuelve aún más complicado si consideramos el alto nivel de automatización en la gestión de la información, algo muy común en muchas empresas e instituciones. La creciente implementación de tecnologías capaces de operar de manera autónoma está transformando a las empresas e instituciones en organizaciones heterogéneas, dónde individuos y dispositivos tecnológicos se combinan y colaboran en la gestión del flujo de información, realizando actividades conjuntamente. En estas organizaciones heterogéneas, la producción, gestión, preservación y acceso a la información son procesos de importancia crítica, por lo que las implicaciones éticas de la transparencia de la información se vuelven aún más desafiantes (Turilli, Floridi, 2009, pp. 105-112).

Actualmente no existe un consenso sobre el tipo de escrutinio necesario, si diferentes áreas afectadas por las computadoras requieren diversas soluciones y si el software, otros factores, o ambos, son la causa de los problemas reclamados. Estas dificultades surgen tanto en un contexto público como en uno privado.

Explorado hasta ahora el panorama que se halla tras la gran cuestión de la eticidad de los algoritmos y la necesidad de establecer un sistema global basado en principio éticos verificables a través de la transparencia, ahora se considera oportuno focalizarnos en el gran enemigo de esta última: la formación de los sesgos.

El término «sesgo», como sugiere un estudio realizado por la *Scientific Foresight Unit (STOA) del EPRS-European Parliamentary Research Service* (Servicio de Investigación Parlamentaria Europeo), es realmente difícil de conceptualizar. En general, hay una doble tendencia en este campo, según el aspecto que más interesa ilustrar. Podríamos definir el sesgo como «una tendencia a preferir a una persona o cosa a otra, y a favorecer a esa persona o cosa», comparándolo al término «prejuicio».

Diferentemente, si analizamos el término que se menciona en estadística, el significado que se le atribuye coincide con el mismo ideado por la Real Academia Española, la cual determina que el sesgo es un «error sistemático en el que se puede incurrir cuando al hacer muestreos o ensayos, se seleccionan o favorecen unas respuestas a otras». En este caso, la acepción es más general y se refiere a cualquier tipo de error sistemático o desviación relevadle tras análisis estadísticos. La diferencia con la primera definición es

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

que esta última no expresa ningún tipo de carga ética problemática, sino simplemente que la introducción de sesgos (estadísticos) en un conjunto de datos puede representar la solución al prejuicio ya existente en una base informativa que refleja con precisión la realidad. De aquí, todas las consideraciones normativas, sea tanto en un contexto nacional que internacional, respecto a la elaboración de soluciones que puedan remediar y eliminar los problemas relacionados con la discriminación como tal y con la calidad de los datos (European Parlamient, 2022).

Interesante es la visión que el investigador Barrio Andrés tiene respecto a las funciones de un algoritmo: «se sabe que los sistemas algorítmicos plantean diversas cuestiones relacionadas con la parcialidad, la injusticia y la discriminación en las decisiones que adoptan, así como con la transparencia, la explicabilidad y la rendición de cuentas en lo que respecta a su funcionamiento o la protección de los datos, la privacidad y otras cuestiones de derechos fundamentales, entre otras. Incluso podemos hablar de "absurdos algorítmicos" para calificar sus cálculos incorrectos» (Barrio Andrés, 2020, pp. 1-6).

La frase proporcionada resalta la importante cuestión sobre la evolución de los sistemas algorítmicos y los retos a los que nos enfrentamos. En otras palabras, si por un lado estos sistemas son sinónimos de mejora de la eficiencia y de la efectividad en la toma de decisiones; por el otro lado, muy a menudo se enfrentan a críticas por la incorporación de fenómenos de discriminación y de prejuicios, muchas veces de forma desapercibida.

De hecho, el problema más grave es cuando los sesgos algorítmicos no solo perpetúan desigualdades existentes, sino que también llegan a amplificarlas. Este fenómeno se genera cuando los algoritmos son alimentados con datos históricos, que bien reflejan las desigualdades de la sociedad. El mito de que la ciencia es un método objetivo en la búsqueda de la verdad declina frente a la conciencia de que los productos científico-tecnológicos están diseñados e ideados por quienes representan solo una parte de la población mundial, aunque sean utilizados por todo el planeta. Este aspecto, junto con la concentración del poder en manos de los lobbies que determinan el destino del mundo, provoca problemas de marginación y exclusión sistemática de una gran parte de la humanidad. La investigadora del MIT, Joy Buolamwini, concluye uno de sus trabajos con una frase que refleja esta idea: «Acknowledging social, cultural, and historic turbulence will be necessary if artificial intelligence is ever to ascend to the elusive stratosphere of fairness and inclusion» (Buolamwini, 2017, pp. 1-116).

Un ejemplo de lesión al derecho de la igualdad y la no discriminación ha surgido en el contexto de la asignación de créditos: los modelos entrenados con datos que reflejan la brecha salarial de género asignan límites de crédito más bajos a las mujeres, perpetuando así la discriminación económica. Otro episodio notable se encuentra en los traductores automáticos como *Google Translate*, donde se descubrió que el sistema asignaba, según su propia lógica, el género que consideraba más adecuado al traducir palabras neutras en un idioma, pero con género determinado en otro, perpetuando algún tipo de discriminación. Por ejemplo, la palabra nurse en inglés puede referirse tanto a un hombre como a una mujer, pero *Google Translate* frecuentemente la traduce como «enfermera», asignándole el género femenino. Del mismo modo, la palabra doctor en inglés se traducía predominantemente como «doctor», en

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

su acepción masculina. Esta tendencia probablemente se debe a que los textos utilizados para entrenar el modelo contenían una mayor probabilidad de encontrar nurse traducida como «enfermera» y doctor traducido en términos masculinos (Stanovsky, Smith, et al., 2019, pp. 1679–1684).

La generación de los sesgos también puede ser causada por las mismas operaciones de cálculos de los algoritmos, cuyo funcionamiento y potencialidad interno no siempre tiene visibilidad. En este sentido se distinguen entre algoritmos «no predictivos» y algoritmos «predictivos». Los primeros se utilizan para llevar a cabo tareas concretas. Su naturaleza puede ser más sencilla o compleja, pero independientemente del tipo de estructura y del tipo de función que se les aplica, los algoritmos no predictivos no sustituyen la norma, sino que la traducen para facilitar su aplicación. De lo contrario, los algoritmos predictivos son los que más preocupaciones generan. La función que se les atribuye es la predictibilidad de determinados resultados a través del manejo de los *big data*, que constan en grandes cantidades de datos a los que el algoritmo tiene acceso y representan la base de partida para extraer las correlaciones necesarias con el fin de realizar predicciones de hechos futuros. En estos casos, las predicciones no podrán ser verificadas, originándose los primeros problemas jurídicos, que bien podrían agravarse si se considera que los algoritmos pueden incluso realizar «aprendizaje no supervisado». En esta ocasión el algoritmo está programado para el autoaprendizaje no controlado, dando vida al notorio efecto de la «caja negra» (fenómeno denominado con el término inglés «*black box*»), dónde ni siquiera el programador sabe precisamente lo que sucede, sintiéndose impotente a la hora de aclarar las razones por las cuales el algoritmo llegó exactamente a ese razonamiento o resultado (Noguerol Díaz, 2020, pp. 265-266).

Por esta razón los principios de transparencia y de explicabilidad recubren un papel fundamental para abordar todas estas cuestiones. Es inviable corregir errores y determinar la responsabilidad de los sujetos o de las cosas que han sido involucradas en ese proceso sin el respeto de estos dos principios, sobre todo cuando los derechos humanos fundamentales afectados son el derecho a la privacidad y a la protección de datos personales. La noción «absurdos algorítmicos» hace resaltar el tema central de este apartado, ósea, que un sistema algorítmico puede cometer errores evidentes a partir de fallos tanto en su diseño como en su aplicación práctica y que claramente son inexplicables.

III. El problema de los sesgos algorítmicos

Luciano Floridi, en una de sus ilustres obras, analiza este aspecto argumentando que es importante investigar la razón por la cual los sesgos, inherentes a nuestros valores, creencias, normas y culturas, también están presentes en los algoritmos utilizados en la inteligencia artificial (IA). Esto se debe, probablemente, a su inclusión en los datos utilizados, al resultado de un entrenamiento inadecuado de los algoritmos, o incluso a su introducción intencionada por parte de los programadores. Tanto los conjuntos de datos como las decisiones tomadas por las computadoras presentan una visión imperfecta del mundo, reflejando los juicios y perspectivas humanas.

Los sesgos en los algoritmos de IA no surgen espontáneamente; más bien, los algoritmos pueden replicar los sesgos existentes si no se manejan adecuadamente en varias

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

etapas clave, incluyendo la recolección de datos (que pueden contener prejuicios preexistentes), la preparación de los datos para el entrenamiento (en la selección y procesamiento de atributos para el algoritmo), y la toma de decisiones durante el desarrollo del sistema inteligente. A partir de esta descripción, el lector podría inferir que los sesgos generados en los softwares de IA son causados exclusivamente por los algoritmos, pero esta idea no es completamente correcta.

La profesora Nuria Belloso Martín se ha encargado de delimitar esta creencia, subrayando que el error más común cometido por quienes no están familiarizados con la mecánica subyacente de la IA es responsabilizar al sistema por los errores cometidos, asumiendo que se trata de una malinterpretación de las instrucciones. Sin embargo, el problema generalmente radica en aspectos como un diseño defectuoso del algoritmo, una selección inadecuada de datos para su entrenamiento o una interpretación incorrecta de los resultados obtenidos.

Más allá de estos factores, el problema también puede derivar de los datos en los que se basan los algoritmos para llevar a cabo sus tareas. A este respecto, el profesor Llano Alonso subraya que, si bien «las decisiones automatizadas son perfectibles (…) y también falibles, en la medida en que no están exentas del factor error ni libres de sesgos, (…) la posibilidad de fallar en el cálculo matemático y la predicción basada en la correlación de datos es menor en la máquina que en el operador humano. Del mismo modo, la presencia de sesgos a lo largo del proceso de cálculo y la resolución del problema es técnicamente menor que el nivel de prejuicios con el que, a veces de manera inconsciente, suele razonar y juzgar el ser humano» (Llano Alonso, 2024, pp. 186–187).

Lo que se busca aclarar es que los sesgos están igualmente presentes tanto en la programación de los algoritmos como en los datos utilizados en los softwares de inteligencia artificial (IA). Según el Decano de la Facultad de Derecho de la Universidad de Sevilla, existe una mayor posibilidad de que la formación de los *bias* provenga principalmente del contenido de los datos, ya que estos proyectan y reflejan los estereotipos y cánones de un contexto político, social y cultural determinado.

Agata Cecilia Amato Mangiameli, destacada académica en filosofía del derecho coincide con esta idea, al considerar que «los datos no son objetivos y los modelos estadísticos representan la realidad modificándola, esto es, orientando sus comportamientos» (Amato Mangiameli, 2019, pp. 107–124). De hecho, otros autores han demostrado que «los datos siempre son ya activos y nunca neutrales» (Iliadis & Russo, 2016, p. 1), y que los algoritmos, según diversos estudios empíricos, reproducen «las disparidades [sobre todo] raciales y de género a través de las personas que los elaboraron o mediante los datos utilizados para testarlos» (Obermeyer, Powers, et al., 2019, p. 447).

Estos investigadores subrayan que no siempre las evidencias producidas por los algoritmos son realmente fiables o justifican elecciones atendibles, debido a la presencia de posibles *bias*, como:

1. Pre-existing bias: sesgos incorporados en el software porque quienes determinan sus contenidos ya poseen sesgos previos.
2. Technical bias: sesgos que surgen de decisiones técnicas o restricciones dentro del sistema.

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

3. Emergent bias: sesgos que provienen de nuevos datos imprevistos introducidos después de la implementación del sistema.
4. Measurement bias: sesgos relacionados con la recolección de los datos utilizados para entrenar la máquina (Obermeyer, Powers, et al., 2019, pp. 447–453).

Las injusticias, por lo tanto, pueden anclarse también en una selección incorrecta de datos de entrenamiento, que debería realizarse de manera cuidadosa y meticulosa (Belloso Martín, 2022, pp. 47–50).

Karen Hao, periodista especializada en el impacto de la IA en la sociedad, presenta un análisis claro sobre los momentos en los que se produce el sesgo algorítmico y la dificultad de mitigarlo. Ella señala que es limitado pensar que la responsabilidad por los prejuicios algorítmicos recae exclusivamente en los algoritmos, al subrayar que «el sesgo puede aparecer mucho antes de que los datos se recopilen y también en muchas otras etapas del proceso de aprendizaje profundo» (Hao, 2019).

Hao identifica tres etapas clave en las que puede originarse este fenómeno: la definición del problema, la recogida de datos y la preparación de datos. Durante la primera fase, cuando los desarrolladores de software abordan la creación de un modelo de aprendizaje profundo, el paso inicial es definir claramente el objetivo del modelo. Por ejemplo, una empresa de tarjetas de crédito que busca predecir la solvencia de sus clientes debe concretar qué significa «solvencia» en términos operativos, lo cual podría estar orientado a aumentar las ganancias o a asegurar la devolución de los préstamos. Según Solón Barocas, profesor asistente en la Universidad de Cornell y experto en equidad en el aprendizaje automático, estas definiciones suelen basarse en intereses comerciales que podrían no considerar cuestiones de justicia o discriminación. Por lo tanto, si un algoritmo considera que otorgar préstamos de alto riesgo maximiza las ganancias, podría actuar de manera depredadora sin que fuera esa la intención original de la empresa.

En la etapa de recolección de datos, los sesgos pueden surgir principalmente de dos formas: porque los datos no reflejan fielmente la realidad (Fernández, 2019, p. 5)[8] o porque replican prejuicios ya existentes.

Un ejemplo claro de los prejuicios en sistemas de inteligencia artificial es el reconocimiento facial. Un sistema entrenado con más imágenes de personas de piel clara que de piel oscura tenderá a funcionar peor al identificar a estas últimas. Un caso notorio fue el de Amazon, cuya herramienta de reclutamiento excluía a candidatas mujeres porque el

8. Ana Fernández indica que los sesgos no intencionados en la inteligencia artificial pueden originarse en diversas etapas del desarrollo del algoritmo, como la recopilación de datos, la metodología de entrenamiento, o las modificaciones realizadas por los programadores. En esencia, para que los algoritmos funcionen de manera efectiva y justa, necesitan ser entrenados con un volumen sustancial de datos que sean de alta calidad y reflejen la diversidad de la población general. Si esto no se logra, existe el riesgo de que los sesgos presentes en los datos de entrenamiento se integren en el algoritmo y se perpetúen, impidiendo el progreso hacia una igualdad de oportunidades. Esto se evidencia en situaciones como la selección de personal donde, si los datos históricos favorecen a un grupo sobre otro, el algoritmo podría replicar esta preferencia. Asimismo, un algoritmo de reconocimiento facial entrenado principalmente con imágenes de hombres tendrá dificultades para identificar correctamente rostros femeninos.

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

algoritmo había sido entrenado con datos históricos que favorecían a los hombres. Este ejemplo ilustra cómo los datos utilizados como fuente primaria para alimentar los algoritmos reflejan la sociedad, marcada por su historia, cultura y, obviamente, prejuicios.

El escritor estadounidense Elwyn Brooks, reconocido por sus libros infantiles, afirmaba que la objetividad no existe: «nunca he visto un escrito, político o no político, que sea imparcial. Todo escrito sigue la tendencia que tiene el escritor, nadie es totalmente objetivo» (Mullane, 2019). Esta falta de objetividad y de imparcialidad siempre ha caracterizado al entorno humano. Las discriminaciones y sesgos suelen influir en las decisiones y acciones humanas y, al alimentar los softwares de inteligencia artificial con estos datos, es ingenuo pensar que el espacio digital sea neutro o que las relaciones generadas por estos sistemas sean igualitarias y equitativas.

Durante la preparación de los datos, también se pueden introducir sesgos al seleccionar los atributos que el algoritmo considerará. Aunque esta fase es distinta de la definición del problema, la selección de atributos es crítica y representa el «arte» del aprendizaje profundo. Por ejemplo, en el contexto de la solvencia crediticia, atributos como la edad, los ingresos o la cantidad de préstamos pagados, y en herramientas de reclutamiento, atributos como el género y la experiencia, tienen un impacto significativo en la precisión de las predicciones. Sin embargo, mientras que la precisión es relativamente fácil de medir, evaluar los sesgos incorporados en el modelo resulta mucho más desafiante.

Además, los sesgos indeseables en la IA pueden surgir según cómo se etiqueten los datos durante el entrenamiento o según cómo el algoritmo se adapte y cambie al recibir nueva información. Para comprender el alcance de este fenómeno, es necesario partir de la función y la tipología de los algoritmos. Estos transforman datos en pruebas que sustentan un resultado, el cual motiva una acción con repercusiones éticas. Los algoritmos de aprendizaje automático, en particular, complican la posibilidad de atribuir responsabilidad por los efectos de las acciones que desencadenan.

El profesor Luciano Floridi plantea que los mecanismos de *deep learning* generan seis cuestiones divididas en dos grupos: epistémicas y normativas. Ambas convergen en un único problema: la dificultad de trazar la cadena de acontecimientos y factores que conducen a un resultado específico. Según Floridi, aunque los algoritmos se utilizan para el bien social, el debate sobre los principios y criterios que deben guiar su diseño y gobernanza es cada vez más relevante. Este autor resalta la necesidad de análisis éticos profundos para mitigar los riesgos asociados con estas tecnologías.

Respecto a la equidad algorítmica, Veale, Binns, Katell y otros coautores plantean dos enfoques principales. El primero propone la intervención de una tercera parte, distinta de aquella que provee los algoritmos, que posea datos sobre características sensibles o protegidas con el fin de identificar y reducir las discriminaciones causadas por los datos y los modelos. El segundo enfoque sugiere un método colaborativo basado en el conocimiento comunitario, empleando recursos de datos generados por la comunidad que incluyen experiencias prácticas en sistemas de aprendizaje profundo.

Diferente es la cuestión de la trazabilidad, que conlleva la dificultad de establecer la responsabilidad moral. Según la recomendación de un profesor oxoniense, la responsabilidad moral debe atribuirse «de manera predeterminada y reversible» a todos los agentes morales que sean causalmente relevantes para una acción específica dentro

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

de una red. Este enfoque se basa en la retropropagación de la teoría de las redes de responsabilidad objetiva en derecho y en el conocimiento común de la lógica epistémica. Luciano Floridi argumenta que dicho enfoque mejora el comportamiento ético de los agentes en la red, de forma similar al principio de responsabilidad objetiva en el ámbito jurídico (Floridi, 2019, pp. 281–293).

4. Transparencia y explicabilidad: Elementos clave para la confianza en la IA

Aclarada la definición de sesgo algorítmico y las fases en las que estos pueden originarse, es necesario abordar el concepto de algoritmo ético. Un algoritmo es un conjunto ordenado y finito de operaciones que transforman entradas en salidas para resolver problemas en cualquier ámbito (Cormen et al., 2009, pp. 5–7) [9]. Tal como la célula es la unidad fundamental de la vida, los algoritmos son esenciales en cualquier software basado en aprendizaje automático o redes neuronales, precedidos en importancia solo por los datos de entrenamiento. Sin estos datos, los algoritmos no podrían funcionar, y estos, a su vez, transportan valores humanos, entre ellos desigualdades y discriminaciones inherentes al contexto social y cultural.

Admitir que los algoritmos son neutrales, ignorando el determinismo tecnológico subyacente, sería contraproducente. Este error no solo subestima el impacto de la tecnología en el desarrollo social, sino que también revela una falta de conciencia sobre su influencia. Un caso emblemático son los algoritmos sociales utilizados por influencers, cuyos contenidos afectan a las comunidades sin que ellos mismos comprendan las implicaciones éticas detrás de ciertas decisiones, como qué contenido priorizar o censurar según las prioridades del momento (Epsilon Tecnología, s.f., párr. 3)[10].

El poder de influencia de los influencers, guiado por las normas de la mercadotecnia influyente, exige reflexionar sobre la falta de transparencia de los algoritmos que utilizan. Según Adrián Todolí y Luminița Pătraș, el artículo 22 del Reglamento General de Protección de Datos (RGPD) establece que los usuarios de algoritmos que afecten derechos deben informar sobre la lógica de funcionamiento del algoritmo. Sin embargo, los autores concluyen que existe poca transparencia en este aspecto. Por ello, proponen

9. El informático estadunidense y profesor de la Dartmouth College Thomas H. Cormen junto con Charles Leiserson, Ron Rivest, e Cliff Stein definían un algoritmo como sigue: «Informally, an algorithm is any well-defined computational procedure that takes some value, or set of values, as input and produces some value or set of values, as output in a finite amount of time. An algorithm is thus a sequence of computational steps that transform the input into the output. You can also view an algorithm as a tool for solving a well-specified computational problem. The statement of the problem specifies in general terms the desired input/output relationship for problem instances, typically of arbitrarily large size. The algorithm describes a specific computational procedure for achieving that input/output relationship for all problem instances. As an example, suppose that you need to sort a sequence of numbers into monotonically increasing order. This problem arises frequently in practice and provides fertile ground for introducing many standard design techniques and analysis tools».

10. Un análisis interesante de las características de los algoritmos de las principales redes sociales de 2023 fue realizado por Épsilon Technologies, una empresa especializada en la analítica avanzada de datos y marketing digital. Por cada algoritmo social, por ejemplo, el algoritmo de Instagram señala el público que atrae y la información que se decide mostrar o priorizar.

IUS ET SCIENTIA • 2024
Vol. 10 • Nº 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

que las autoridades administrativas aumenten el control sobre la normativa, permitiendo que tanto los creadores de contenido como el público entiendan mejor cómo operan los algoritmos que afectan sus vidas (Pătraș & Todolí, 2022, p. 54).

El riesgo real es considerar a los algoritmos como entes neutrales e incuestionables, divinizándolos. Esto podría llevar a los seres humanos a convertirse en sujetos subordinados a los caprichos de la tecnología (Martin, 2019, pp. 835–850). Si bien algunos temen que los seres humanos puedan convertirse en marionetas de los algoritmos, también es crucial reflexionar sobre cómo prevenir, desde un punto de vista técnico, la introducción de prejuicios por parte de quienes diseñan y procesan los datos.

Una posible solución es desarrollar una inteligencia artificial débil que no limite la libertad humana, sino que recopile datos y ofrezca asistencia moral a los agentes en la toma de decisiones. Sin embargo, este enfoque enfrenta desafíos significativos, ya que el sistema debería estar entrenado con información sobre los valores y principios del agente individual, así como sobre sesgos cognitivos comunes que afectan las decisiones morales. Este sistema debería advertir a los agentes sobre factores físicos y ambientales que puedan influir en sus juicios, guiándolos hacia acciones que consideren éticas según sus propios valores.

La función de este software, como observador constante de un determinado agente y su desarrollo ético, sería alertar al sujeto frente a la posibilidad de una eventual amenaza o desviación de su propia moralidad, guiándolo hacia la integridad moral individual. Este sistema actuaría como un entrenador moral permanente que, frente a cada reto, determinaría un objetivo moral y los pasos necesarios para alcanzarlo.

Si trasladamos este concepto al análisis de sesgos en los algoritmos, es decir, los prejuicios introducidos tanto por los diseñadores como por las etiquetas de datos, podríamos vislumbrar una tecnología capaz de reducir *bias* y eliminar influencias externas que generen discriminación o resultados injustos (Savulescu & Maslen, 2015, pp. 79–95). Sin embargo, las ideas de esta escuela de pensamiento generan dudas inquietantes. Aunque el autor asegura un control pleno, esta propuesta podría ser peligrosa para la humanidad, marcando un antes y un después en la percepción de la autonomía moral individual.

Según Julian Savulescu, ser autónomos significa «vivir mi vida de acuerdo con lo que yo pienso que es bueno, no según aquello que los demás consideran bueno» (Savulescu, 2012, p. 169), siempre que las decisiones estén justificadas con razones normativas sólidas. La autonomía implica la capacidad de tomar decisiones únicas, pero la introducción de una IA moral podría limitar este desarrollo personal. Dado que el ser humano evoluciona constantemente a través del contacto social, ambiental y las experiencias empíricas, resulta difícil creer que una IA moral, actuando como un entrenador rígido, pueda guiar eficazmente sin infringir la libertad individual y la subjetividad de cada persona.

Esta limitación entra en conflicto con el artículo 10.1 de la Constitución Española, que protege la dignidad y los derechos fundamentales, así como con los artículos 1, 3, 6, 8 y 21 de la Carta de Derechos Fundamentales de la Unión Europea y los artículos 8, 9 y 10 del Convenio Europeo de Derechos Humanos. Además, la monitorización constante de los valores y principios de un individuo podría exacerbar el problema de la vigilancia masiva, incrementando los riesgos de abuso por parte de criminales o expertos en IA con intenciones cuestionables.

IUS ET SCIENTIA • 2024
Vol. 10 • Nº 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

Otro aspecto preocupante es la creciente brecha entre los tecno-ricos y los tecno-pobres, según el profesor Pérez Luno. La falta de acceso a una IA moral podría marginar aún más a las clases vulnerables, cuyos pensamientos podrían ser ignorados y sus necesidades, aisladas. Este tipo de exclusión social ya es frecuente en la actualidad, pero la tecnología podría amplificarla.

Savulescu también plantea que una IA moral podría no solo guiarnos, sino llevarnos a un nivel superior de cognición y altruismo. Sin embargo, reconoce los límites de esta visión: «nuestro conocimiento sobre estas materias es muy limitado. […] No parece probable que podamos efectuar mejoras significativas en la moralidad antes de que algunos individuos moralmente pervertidos hagan un mal uso de nuestro conocimiento científico y tecnología, con consecuencias fatales» (Savulescu, 2012, p. 234). Esta incertidumbre también dificulta la asignación de responsabilidad por decisiones tomadas por algoritmos que violen derechos humanos fundamentales, ya que los algoritmos carecen de subjetividad jurídica.

En este sentido, la responsabilidad debería recaer sobre quienes diseñan, implementan y utilizan estos sistemas. Según Kraemer, Van Overveld y otros, al crear algoritmos, los desarrolladores toman posiciones éticas implícitas al decidir qué es bueno o deseable: «expressing a view on how things ought to be or not to be» (Kraemer et al., 2020, pp. 251–260).

Un movimiento filosófico que aborda este dilema es el de la «objetividad mecánica» (Daston & Galison, 1992, pp. 81–128), que considera los algoritmos más objetivos que las perspectivas humanas, ya que, supuestamente, no están influenciados por valores o intereses subjetivos. Sin embargo, Florian Pethig y Julia Kroenung identificaron un fenómeno interesante: un grupo de mujeres percibe las decisiones algorítmicas como más fiables que las tomadas por hombres (Pethig & Kroenung, 2023, pp. 637–652).

Para abordar los retos éticos y técnicos de los algoritmos, Michael Kearns y Aaron Roth proponen en su obra un enfoque integral. Ambos autores destacan que los problemas de imparcialidad y privacidad surgen principalmente de los diseños algorítmicos, especialmente en sistemas de aprendizaje automático. Argumentan que la interacción de algoritmos sencillos con datos complejos genera modelos predictivos igualmente complejos, aumentando el riesgo de resultados sesgados.

Este modelo es el resultado de un aprendizaje automático dotado de un perfil propio, diferente del que fue realizado por el diseñador. Por lo tanto, la solución para «cerciorarse de que los efectos de los modelos cumplen con las normas sociales que se desea respetar es diseñar estos objetivos directamente en nuestros algoritmos» (Kearns & Roth, 2020, pp. 14–23). Las reflexiones de estos dos autores se dividen entre privacidad e imparcialidad, aunque en este texto se considera más oportuno concentrarse en el segundo aspecto.

El ejemplo propuesto por los autores es el caso de los sesgos por analogía. Aunque cayeron en el olvido tras haber sido eliminados de las pruebas de acceso universitario en 2005 en Estados Unidos, un grupo de investigadores, en 2016, decidió darles vida nuevamente. Sometieron el modelo *Word Embedding* (literalmente «encaje de palabras»), disponible de forma gratuita en Google, a una prueba basada en analogías con el fin de calcular valores estadísticos sobre las concurrencias de palabras. La técnica de Google *word2vec* demostró que los sesgos ya estaban presentes en los documentos de

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

origen con los que el algoritmo había sido entrenado. Por ejemplo, el término «dama» se asocia a «pendientes», mientras que «genio» a «sobrino». Este sexismo es evidente, ya que las mujeres se ven asociadas al brillo de los adornos, mientras que los hombres al brillo personal. El tema se vuelve relevante cuando los modelos sesgados, como el *Word Embedding*, actúan como componentes de otras aplicaciones, difundiendo como una mancha de petróleo el problema de los sesgos.

Desde una perspectiva más amplia, el problema radica directamente en los datos de entrenamiento, cuyo uso en modelos de aprendizaje automático implica una expansión de los sesgos ya presentes en dichos datos. Otro aspecto que debe considerarse es la influencia de estos sesgos en relación con el sujeto al que está dirigido el resultado. En este contexto, se discute la diferencia entre los sesgos que afectan factores generales, como el género, la raza o la edad, y aquellos más individuales, como suele ocurrir cuando la predicción se dirige a un sujeto específico que, por ejemplo, desea conocer la probabilidad de ser beneficiario de una beca. Más allá de esta diferencia, lo que se evidencia es que los criterios y las normas humanas siempre deben ocupar un papel central en el debate.

Esta discusión conduce al lector a preguntarse cómo encasillar dentro de una definición el concepto de «imparcialidad» y qué instrumentos usar para alcanzarla. En general, cuando hablamos de algoritmos, la imparcialidad suele equipararse a la «paridad estadística».

Así como sugieren los profesores Kearns y Roth, es fundamental analizar el sujeto que debe ser protegido cuando se aborda la temática de los algoritmos y la imparcialidad. En un escenario hipotético, se presenta un planeta poblado por dos razas: los Círculos y los Cuadrados. La preocupación principal radica en delimitar la discriminación hacia los Círculos al aprobar solicitudes de crédito, estableciendo que los Cuadrados sean un grupo protegido. La paridad estadística, en este caso, no determina cuántos créditos se conceden a cada grupo, sino que asegura una proporción similar de aprobación entre ellos.

Esta paridad puede entenderse como una subcategoría defectuosa de la imparcialidad. Una objeción frecuente es la ineficacia del algoritmo basado en paridad estadística, ya que puede ignorar aspectos individuales relevantes y dejar parte de la decisión al azar. Sin embargo, esta crítica no es tan grave si se comprende que la paridad estadística actúa como una restricción y no como un objetivo de predicción. Un algoritmo deficiente, que cumple con la paridad estadística, no implica la inexistencia de algoritmos eficaces que aprueben los créditos adecuadamente para los individuos «correctos» de ambos grupos. El propósito del algoritmo puede ser minimizar el error de predicción o aumentar los beneficios, siempre bajo la limitación de que los créditos cumplan con la equivalencia entre razas. En este sentido, la concesión aleatoria de créditos puede ser tranquilizadora, ya que demuestra que la definición de imparcialidad puede lograrse.

Además, otorgar préstamos aleatorios puede ser una estrategia válida para cumplir con la paridad estadística mientras se recopilan datos. Si un nuevo prestamista no tiene información sobre la relación entre los atributos del solicitante y la devolución del préstamo, puede conceder créditos de manera aleatoria hasta contar con datos suficientes para tomar decisiones más informadas, manteniendo la imparcialidad inicial. En

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

aprendizaje automático, esta práctica se conoce como «exploración», un período en el cual se recopilan datos en lugar de tomar decisiones óptimas. También pueden existir escenarios donde la ceguera deliberada de las decisiones aleatorias sea conveniente, como al distribuir un número limitado de entradas gratuitas a un evento público sin que existan candidatos más cualificados que otros.

Una objeción más grave es que la paridad estadística no considera el valor crediticio final de cada solicitante. Si las tasas de devolución difieren entre grupos, mantener la paridad puede llevar a decisiones difíciles y potencialmente injustas. Por ejemplo, cumplir con la paridad estadística otorgando créditos a un porcentaje fijo de solicitantes solventes de ambos grupos puede ser injusto para aquellos que, aunque solventes, se ven rechazados debido a la limitación impuesta. Además, prestar a quienes devolverán el préstamo es financieramente beneficioso, mientras que prestar a quienes no lo devolverán puede generar pérdidas.

5. Compromisos entre imparcialidad y exactitud: Un desafío técnico y moral

La paridad estadística no entra en conflicto con la exploración de datos, pero sí con la «explotación óptima de decisiones». En estos casos, no se puede simplemente optimizar la exactitud; se debe maximizar dentro de los límites de la paridad estadística. Esto puede llevar a soluciones insatisfactorias, como negar créditos a solicitantes solventes o concederlos a quienes no los devolverán. La sociedad debe aceptar compromisos intermedios entre imparcialidad y precisión en los modelos, tomando decisiones basadas errores de predicción, asegurando que las tasas de falsos rechazos sean similares para diferentes grupos, introduciendo el concepto de «igualdad de falsos negativos». Aunque esto puede no ser satisfactorio para individuos específicos, aborda la imparcialidad a nivel de grupo. Optimizar la exactitud predictiva en múltiples poblaciones tiende a favorecer a la mayoría, generando disparidades en los «falsos rechazos». La única respuesta sensata desde una perspectiva científica, reguladora y moral es reconocer estas tensiones y gestionar directamente los compromisos entre «exactitud» e «imparcialidad».

La exploración cuantitativa y sistemática de soluciones de compromiso entre exactitud e imparcialidad es crucial. En el contexto de la Universidad de la Santa Imparcialidad, el aprendizaje automático busca minimizar errores en las predicciones sin restricciones de imparcialidad. Este proceso implica encontrar el valor de la nota de corte que minimice el número total de errores (rechazos de estudiantes exitosos y aceptación de fracasados) sin considerar la raza. De igual forma, se podría buscar un modelo que minimice la parcialidad global. Esto implica calcular la «puntuación de parcialidad» de un modelo mediante la magnitud de la diferencia entre el número de falsos rechazos en los grupos de Círculos y Cuadrados. Aplicando principios de aprendizaje automático, se pueden diseñar algoritmos que minimicen la parcialidad. Los modelos se evalúan según dos criterios: errores cometidos y valor de parcialidad, permitiendo seleccionar la mejor solución de compromiso.

A veces, existen modelos que son claramente inferiores: desplazar, por ejemplo, la línea de corte óptimo hacia la izquierda, aceptando a tres estudiantes destinados al

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

fracaso, incrementa los errores sin mejorar la imparcialidad. Este modelo sería inferior al óptimo en términos de errores. La frontera de Pareto, que conecta los modelos no dominados, representa las soluciones «razonables» que equilibran exactitud e imparcialidad. Cualquier modelo fuera de esta frontera es inferior y debe ser eliminado del análisis. La frontera de Pareto cuantifica las soluciones de compromiso entre exactitud e imparcialidad. Sin embargo, no indica cuál es la mejor opción, ya que esto depende de la importancia relativa de cada criterio. Algoritmos prácticos para el aprendizaje automático pueden trazar esta frontera, aunque son más complejos que los estándares. Un enfoque sería inventar un objetivo numérico que combine error y parcialidad, permitiendo identificar puntos en la frontera de Pareto al modificar las ponderaciones.

A pesar de la incomodidad que pueda generar considerar soluciones de compromiso cuantitativas entre exactitud e imparcialidad, esta práctica es inevitable. Una vez elegido un modelo de toma de decisiones, solo existen dos posibilidades: el modelo no está en la frontera de Pareto (y es inferior) o sí aparece en ella, implicando una ponderación implícita de la importancia del error y la imparcialidad. Pensar en términos menos cuantitativos no cambia esta realidad, solo la oculta.

Convertir la solución de compromiso en un análisis cuantitativo no elimina la importancia del juicio humano, la política y la ética, sino que se enfoca en decidir qué modelo de la frontera de Pareto es el mejor. Decisiones basadas en factores no cuantificables, como el propósito social de proteger a un grupo, son esenciales. Por ejemplo, los sesgos raciales en anuncios publicitarios y en decisiones de crédito tienen impactos diferentes. Es preferible insistir en la igualdad de tasas de falsos rechazos entre grupos raciales, incluso a costa de reducir beneficios bancarios, para evitar concentrar errores en un grupo racial específico. Este es el compromiso necesario para lograr garantías sólidas de imparcialidad.

Antes de abordar el papel del juicio humano en la elección de un modelo en la frontera de Pareto, debemos resolver la cuestión de qué concepto de imparcialidad utilizar. Contamos con varias alternativas razonables. La paridad estadística podría ser adecuada en escenarios donde solo se desee distribuir oportunidades equitativamente, como en la asignación de entradas gratuitas para un concierto, sin cuestiones de mérito implicadas. En decisiones crediticias, la igualdad aproximada de falsos negativos entre grupos sería más adecuada. Para auditorías fiscales, la igualdad de falsos positivos sería crucial, ya que los falsos positivos producen perjuicios significativos.

Aspirar a modelos lo más exactos e imparcial posible es natural, pero también debemos definir la imparcialidad de manera robusta. Sin embargo, existen combinaciones de criterios de imparcialidad que no pueden lograrse simultáneamente, incluso ignorando la exactitud. Por ejemplo, la igualdad de falsos positivos y negativos junto con la igualdad de valor predictivo de positivos son tres definiciones de imparcialidad que, aunque razonables por separado, son imposibles de conseguir juntas. Esta realidad resalta la necesidad de compromisos entre diferentes nociones de imparcialidad. Las restricciones matemáticas sobre este aspecto destacan que, aunque los algoritmos pueden calcular la frontera de Pareto, no pueden decidir qué definición de imparcialidad utilizar ni qué modelo elegir. Estas decisiones son subjetivas y normativas, y no pueden ser resueltas únicamente por la ciencia.

Además, una decisión crucial es la elección de los grupos a proteger. Hemos mostrado ejemplos de sesgos por sexo y raza, pero también existen otros factores, como

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

la edad, discapacidad, riqueza, nacionalidad y orientación sexual. La elección de los grupos a proteger es una decisión social y no puede ser determinada por algoritmos.

Un fenómeno reciente es la «manipulación de la imparcialidad», donde se protegen varios grupos solapados a costa de discriminar a intersecciones entre ellos. Por ejemplo, repartir entradas equitativamente entre hombres, mujeres, Círculos y Cuadrados puede resultar en concentrar entradas en ciertos subgrupos y excluir a otros. Para evitar esto, es necesario definir claramente la protección para subgrupos más específicos.

El aprendizaje automático puede enfrentar la manipulación de la imparcialidad mediante un enfoque de juego, donde, por un lado, un Aprendiz intenta minimizar errores, mientras que, por el otro lado, un Regulador señala continuamente al Aprendiz los subgrupos discriminados por su modelo. Este proceso garantiza un modelo imparcial para todos los subgrupos relevantes, manteniendo la exactitud dentro de lo posible.

Considerar la imparcialidad para subgrupos más estrictos lleva a la lógica de proteger a individuos. Sin embargo, buscar una imparcialidad individual absoluta puede ser impracticable y costosa en términos de exactitud. Encontrar formas razonables de garantizar imparcialidad individual es un área prometedora de investigación. Además, existen preocupaciones en torno a la imparcialidad antes y después de construir los algoritmos. Si los datos de entrenamiento están sesgados, los algoritmos reproducirán y amplificarán estos sesgos. Por ejemplo, decisiones policiales basadas en datos sesgados pueden crear bucles de retroalimentación que perpetúan la discriminación. En definitiva, aunque sea posible diseñar algoritmos imparciales, la implementación efectiva requiere una comprensión profunda de los contextos sociales y la adopción de prácticas limpias en la recopilación de datos.

Si bien al principio del texto se ilustraron los problemas ético-jurídicos en la construcción de un algoritmo, posteriormente se desarrolló el problema de la imparcialidad desde un punto de vista matemático-económico, mostrando qué tipo de soluciones se podrían tomar al respecto. La razón de toda esta explicación radica, principalmente, en la necesidad de que un jurista conozca las definiciones matemáticas que constituyen este complejo objeto de investigación. Solo gracias a la delimitación de estos significados es posible construir una teoría útil desde el punto de vista del derecho (Kearns & Roth, 2020, pp. 85–135).

6. Hacia una regulación ética de los algoritmos en Europa

Volviendo al punto de partida, hemos comprendido el papel que ocupa la ética en la elaboración de los algoritmos y en la recogida de los datos. Esta se presenta como una subcategoría del macroconcepto de «ética», que fundamentalmente consiste, según Sara Degli Esposti, en la «formulación de juicios morales» (Degli Esposti, 2023, p. 34). Aunque algunos autores, como Luciano Floridi, no están de acuerdo con la definición de «ética de la IA» (Floridi, 1999, pp. 33–52), una verdad indiscutible es que esta ética se expresa a través de principios y directrices que guían el diseño, el despliegue o la adopción de la IA. Las diversas formas de injusticia algorítmica derivan de clasificaciones arbitrarias, la validación de sesgos, estereotipos, distorsiones y exclusiones ya presentes o introducidas en un conjunto de datos, o bien de la cristalización de discriminaciones estructurales. Estas tareas están caracterizadas por un

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

estatus ontológico y moral, relacionado con el hecho de que quienes generan la IA somos nosotros mismos.

Por lo tanto, es necesario estudiar cuestiones ontológicas y tecnocientíficas, ya que este tema se está convirtiendo en uno de los principales tópicos en el ámbito de la investigación científica. Los procesos selectivos basados en algoritmos, además, están revestidos de una opacidad que afecta tanto al proceso lógico de producción de resultados como a las «componentes preceptivas» de las normas, sujetas a «constantes evoluciones debidas a inferencias que producen un conocimiento probable pero incierto» (Vantin, 2021, pp. 195–197). La transparencia, por lo tanto, ocupa un papel central en el estudio de los algoritmos «éticos».

En la dictadura de la probabilidad, definida por la toma de decisiones realizada por algoritmos, las brechas existentes en la realidad conllevan la exclusión de segmentos enteros de la población mundial. Este dato, trasladado a una existencia virtual, implica un rediseño de la sociedad y una influencia en el modo de conocer y percibir el mundo, incluyendo a nosotros mismos, de forma completamente imparcial (Floridi, 2017, pp. 133–134). La falta de transparencia supone un gran riesgo, especialmente para el derecho antidiscriminatorio.

Serena Vantin aborda este aspecto destacando el doble efecto del fenómeno: por un lado, se perpetúan estereotipos y prejuicios; por el otro, se solidifica una situación de discriminación estructural basada en un historial de discriminaciones anteriores, que se mantienen y replican en decisiones futuras. Estas funciones, definidas como «función veridictiva» y «función predictiva» (Vantin, 2021, pp. 374–376), cubren las nuevas tecnologías en el derecho antidiscriminatorio.

Se generan de esta manera algunas «formas de desobediencia posibles»: desde la auditoría algorítmica (con la difusión pública de los resultados) hasta el *ofuscamiento* (un conjunto de técnicas orientadas a despistar a los algoritmos alterando conscientemente los datos proporcionados por nuestras interacciones), pasando por la recolección crítica de datos excluidos por los análisis de los softwares, el rediseño (diseño de arquitecturas alternativas) y el *opt-out*, es decir, la práctica de la desconexión. Se necesita urgentemente idear propuestas políticas para mejorar la situación actual. Entre las más relevantes se encuentran las elaboradas en el documento publicado el 25 de julio de 2022 por STOA, titulado *Auditing the quality of data sets used in algorithmic decision-making systems*.

Las opciones propuestas se expresan como sigue:

1. No crear nuevas regulaciones centradas específicamente en los sesgos, sino enfocarse en los desajustes entre las diferentes herramientas regulatorias.
2. Definir un enfoque preventivo, fortaleciendo la mitigación del sesgo desde la recopilación de datos.
3. Crear certificados que puedan servir para garantizar la estandarización de las bases de datos.
4. Otorgar derechos de transparencia a los sujetos del sistema de IA, abriendo una ventana para encontrar la fuente de los resultados sesgados.
5. Facilitar la implementación de la Ley de Inteligencia Artificial.

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

De todas estas propuestas, las únicas que han sido realmente explotadas, sin que la misma agencia europea expresara dudas sobre su viabilidad, son la tercera y la última (Parlamento Europeo, 2022, p. 12)[11].

No es casual que, el 12 de julio de 2024, se haya publicado definitivamente el primer Reglamento de la Inteligencia Artificial (*Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024*), que establece normas armonizadas en materia de IA. A diferencia de los enfoques de América y China, este reglamento está enfocado completamente en la difusión de los principios que rigen una cultura ético-algorítmica y en la importancia de minimizar la formación de sesgos algorítmicos. Aunque su entrada en vigor se realizará el 2 de agosto de 2026, salvo por algunos capítulos que serán aplicables antes, finalmente parece que se ha logrado una colaboración y armonización entre los Estados miembros de la UE.

Un análisis detallado del reglamento permite visualizar el papel que aún ocupa el ser humano en un contexto digitalizado. El artículo 1 de la Ley de Inteligencia Artificial establece lo siguiente:

> El objetivo del presente Reglamento es mejorar el funcionamiento del mercado interior y promover la adopción de una inteligencia artificial (IA) centrada en el ser humano y fiable, garantizando al mismo tiempo un elevado nivel de protección de la salud, la seguridad y los derechos fundamentales consagrados en la Carta, incluidos la democracia, el Estado de Derecho y la protección del medio ambiente, frente a los efectos perjudiciales de los sistemas de IA (en lo sucesivo, 'sistemas de IA') en la Unión, así como prestar apoyo a la innovación.

Más allá de un análisis exhaustivo del texto, lo cual será desarrollado en la segunda parte de este trabajo, es necesario destacar la preocupación del reglamento respecto a la formación de sesgos y las medidas propuestas para enfrentarlos. En particular, la tercera propuesta política de STOA, mencionada anteriormente, ha tomado forma concreta con la inclusión de certificados de calidad para los conjuntos de datos. Esta innovación busca proteger a la sociedad frente a la proliferación de sesgos en el uso de la IA.

El Capítulo III del reglamento, dedicado a los sistemas de IA de alto riesgo, establece en su artículo 11 la obligatoriedad de la documentación técnica que garantice el cumplimiento de los requisitos. Según el reglamento, esta documentación debe ser elaborada antes de la introducción del sistema en el mercado o de su puesta en servicio, y mantenerse actualizada. A tal efecto, se establece un formulario simplificado para pymes y startups, con el objetivo de facilitar el cumplimiento sin comprometer la conformidad.

11. Los límites de cada opción fueron delineados de forma extensa por parte de *Scientific Foresight Unit* (STOA) del *EPRS-European Parliamentary Research Service*. Respecto a la primera opción, la hipótesis de regular en forma específica los sesgos se descartaron por el temor a que una nueva regulación pudiera ser innecesaria y resulte en sub-regulación, considerando también la posibilidad de que las regulaciones actuales puedan ser adecuadas si se modifican o si se añaden directrices claras sobre equidad y sesgos. Con la segunda opción, que bien predica la mitigación del sesgo desde la recopilación de datos trámite un control *ex ante*, no se enfrenta directamente el problema de la falta de normativas claras o estándares comunes. La cuarta idea, que coincide con el reconocimiento de los derechos de transparencia, aparte no encajar con la visión de Luciano Floridi expuesta en las páginas anteriores, tampoco está delineada claramente en la Ley de la IA publicada el 12 de Julio de 2024.

IUS ET SCIENTIA • 2024
Vol. 10 • Nº 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

Asimismo, el artículo 12 estipula la necesidad de registrar estos sistemas en un archivo durante todo su ciclo de vida, garantizando un nivel de trazabilidad adecuado.

Además, las sanciones administrativas, que pueden llegar hasta los 35 millones de euros, junto con la creación de organismos como el Consejo Europeo de Inteligencia Artificial, un Foro Consultivo y un Grupo de Expertos Científicos Independientes, son prueba de la dirección política adoptada por la Unión Europea. Esta dirección, definida en el Considerando (8) del reglamento, promueve un enfoque europeo centrado en el ser humano y comprometido con una IA segura, confiable y ética.

Estas disposiciones son fundamentales para limitar la difusión de sesgos algorítmicos, reconociendo a los Estados miembros la capacidad de ejercer un control efectivo sobre los sesgos existentes y los que podrían surgir en el futuro. En el punto (27) del reglamento, se subraya la importancia de desarrollar códigos de conducta basados en principios éticos y mejores prácticas. Estos principios, aunque no vinculantes, derivan del trabajo del Grupo Independiente de Expertos de Alto Nivel sobre IA, publicado en 2019.

Tomando estos principios como guía, el reglamento pone énfasis en la salvaguarda de la salud, la seguridad y los derechos fundamentales de las personas, así como en la mitigación de sesgos. Esto es especialmente relevante cuando los resultados de los sistemas de IA retroalimentan futuros procesos. Según el reglamento, los sesgos pueden estar presentes en los datos históricos o surgir durante la implementación de los sistemas en contextos reales. Para minimizar estos riesgos, se requiere que los conjuntos de datos sean completos y precisos, reflejando las características específicas del entorno en el que se utilizará el sistema.

Estas medidas buscan garantizar una supervisión humana constante, esencial para prevenir riesgos y reducir daños. Casos como el de Judith Sullivan, ilustran las limitaciones de los algoritmos sin supervisión humana. En este caso, el sistema «NhPredict» no consideró necesidades esenciales de la paciente, como el cuidado de heridas o su incapacidad para subir escaleras, fallando en proporcionar soluciones adecuadas. Este ejemplo demuestra cómo la dependencia excesiva en algoritmos puede tener consecuencias perjudiciales.

En definitiva, el reglamento busca armonizar las normativas entre los países miembros para establecer un marco que regule la creación, desarrollo y adopción de sistemas de IA. Si bien este reglamento parece ser exhaustivo y estar diseñado para promover una IA ética y transparente, aún quedan desafíos futuros por abordar.

Entre las conclusiones más relevantes, se identifican dos aspectos principales: uno político-económico y otro moral-ético. Aunque la Unión Europea es pionera en establecer un marco ético y normativo, el poder de influencia de empresas y organizaciones internacionales fuera de la UE, que optan por autorregularse bajo principios propios, representa un reto significativo. Por ejemplo, los «Principios para la IA – Objetivos para desarrollar una IA beneficiosa» (2023) son una iniciativa que refuerza la autorregulación en otros contextos.

El segundo aspecto tiene que ver con la detección y neutralización de sesgos, necesaria para fomentar la igualdad tanto en el diseño de los algoritmos como en la definición de los conjuntos de datos. El desafío no radica únicamente en la formulación de conceptos como transparencia, equidad o justicia, sino en su implementación efectiva, considerando distinciones clave, como la igualdad sustancial frente a la igualdad formal.

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

Al fin y al cabo, estamos ante conceptos morales que, como el amor, la fraternidad o la justicia, son difíciles de encasillar, especialmente cuando se enfrentan al objetivo principal de promover el desarrollo económico en el ámbito de las tecnologías emergentes.

7. Conclusiones

A lo largo de este trabajo, he intentado desentrañar las múltiples capas de complejidad que envuelven el diseño y uso ético de los algoritmos, en particular, los retos que plantea la IA en términos de equidad, transparencia y responsabilidad. Al reflexionar sobre estos temas, considero que estamos frente a un punto de inflexión, no solo en la evolución tecnológica, sino también en cómo como sociedad afrontamos las implicaciones éticas y jurídicas de nuestras creaciones.

Es innegable que los sesgos algorítmicos no son únicamente problemas técnicos; son reflejos de nuestras propias imperfecciones como seres humanos. Desde los datos que utilizamos hasta las decisiones que tomamos al programar, todo está impregnado de valores, prejuicios e historias que, si no se abordan con rigor, pueden perpetuar desigualdades estructurales. Sin embargo, sería injusto limitar la conversación a los problemas: el potencial de la inteligencia artificial para transformar positivamente nuestras vidas es enorme, siempre que se utilice con prudencia y sentido de justicia.

De igual modo, la transparencia y la explicabilidad no son solo atributos deseables; son la base para construir sistemas que respeten la dignidad humana. Sin embargo, hay que reconocer que estos principios pueden entrar en conflicto con otros objetivos, como la eficiencia o la precisión, no debe llevarnos al conformismo.

Creo firmemente que es posible encontrar equilibrios creativos, aunque ello requiera una colaboración constante entre disciplinas. En este sentido, el Reglamento Europeo sobre la Inteligencia Artificial marca un hito crucial al colocar la ética en el centro del debate regulatorio. No obstante, su implementación será la verdadera prueba de su eficacia, y aquí radica un llamado a la acción para todos los involucrados: académicos, legisladores, desarrolladores y ciudadanos.

Por último, este estudio no pretende ser una conclusión definitiva, sino una invitación al diálogo continuo. Más allá de las normas y los algoritmos, la verdadera solución radica en nuestra capacidad colectiva para redefinir qué entendemos por justicia, equidad y progreso en un mundo cada vez más moldeado por la tecnología.

Así como la ética guía nuestras acciones, es también nuestro deber permitir que oriente las decisiones de las máquinas, sin olvidar nunca que, detrás de cada línea de código, estamos nosotros, los humanos, responsables y beneficiarios de lo que creamos.

Bibliografía

Amato Mangiameli, A. (2019). Algoritmi e big data. Falla carta sulla robotica. *Rivista di Filosofia del Diritto, 1*, pp.107-124.

Barocas, S., Hood, S., & Ziewitz, M. (2013). *Governing algorithms: A provocation piece.* SSRN. http://dx.doi.org/10.2139/ssrn.2245322.

Belloso Martín, N. (2022). La problemática de los sesgos algorítmicos (con especial referencia a los de género): ¿Hacia un derecho a la protección contra los sesgos? En F. H. Llano Alonso

IUS ET SCIENTIA • 2024
Vol. 10 • Nº 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

(Dir.), J. Garrido Martín, & R. D. Valdivia Jiménez (Coords.), *Inteligencia Artificial y Filosofía del Derecho* (pp. 47-50). Laborum Ediciones.

Boix Palop, A. (2020). Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones. *Revista de Derecho Público: Teoría y Método, 1*, pp.223-270.

Bostrom, N., & Yudkowsky, E. (2011). The ethics of artificial intelligence. En W. Ramsey & K. Frankish (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 1-2). Cambridge University Press.

Buolamwini, J. A. (2017). *Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. Massachusetts Institute of Technology.

Cerrillo I Martínez, A. (2019). El impacto de la inteligencia artificial en el derecho administrativo: ¿Nuevos conceptos para nuevas realidades técnicas? *Revista General de Derecho Administrativo, 50*.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms* (3ª ed.). The MIT Press.

Daston, L., & Galison, P. (1992). The image of objectivity. Representations, (40), 81–128. https://doi.org/10.2307/2928741.

Degli-Esposti, S. (2023). *¿Qué sabemos de? La ética de la inteligencia artificial*. Los libros de las Cataratas, pp. 33-34.

Epsilon Tecnología. (s.f.). Algoritmos en redes sociales: ¿Qué importancia tienen? Recuperado de https://epsilontec.com/algoritmos-en-redes-sociales-que-importancia-tienen/.

Fernández, A. (2019). Inteligencia artificial en los servicios financieros. *Boletín Económico. Artículos Analíticos del Banco de España, 2*, p. 5.

Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology, 1*(1), 33–52. DOI:10.1023/A:1010018611096.

Floridi, L. (2017). *La quarta rivoluzione: Come l'infosfera sta cambiando il mondo*. Raffaello Cortina.

Floridi, L. (2019). *Etica dell'intelligenza artificiale: Sviluppi, opportunità, sfide*. Raffaello Cortina Editore.

García-Marzá, D. (2023). Ética digital discursiva: De la explicabilidad a la participación. *Daimon. Revista Internacional de Filosofía, 90*, 99–114.

Gutiérrez David, M. E. (2021). Administraciones inteligentes y acceso al código fuente y los algoritmos públicos: Conjurando riesgos de cajas negras decisionales. *Derecom: Revista Internacional de Derecho de la Comunicación y las Nuevas Tecnologías, 30*, p. 50.

Hao, K. (2019, febrero 8). Cómo se produce el sesgo algorítmico y por qué es tan difícil de entenderlo. MIT Technology Review.

Hill, R. K. (2016). What an algorithm is. *Philosophy & Technology, 29*, 35–59.

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society, 2*(1), 1–12. DOI:10.1177/2053951716674238.

Kearns, M., & Roth, A. (2020). *The ethical algorithm: The science of socially aware Kearns, M., & Roth, A. (2020). The ethical algorithm: The science of socially aware algorithm design (Trad. GEA Textos, S.L.). Wolters Kluwer España.*

Kraemer, F., Van Overveld, K., & Peterson, M. (2020). Is there an ethics of algorithms? *Ethics and Information Technology, 13*(3), 251–260. DOI:10.1007/s10676-010-9233-7.

Llano Alonso, F. (2024). *Homo ex machina: Ética de la inteligencia artificial y derecho digital ante el horizonte de la singularidad tecnológica*. Tirant lo Blanch.

IUS ET SCIENTIA • 2024
Vol. 10 • Nº 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

Barrio Andrés, M. (2020). Retos y desafíos del Estado algorítmico de Derecho. *Real Instituto Elcano. Análisis del Real Instituto Elcano, 82*. https://www.realinstitutoelcano.org/analisis/retos-y-desafios-del-estado-algoritmico-de-derecho/.

Mager, A. (2012). Algorithmic ideology: How capitalist society shapes search engines. Information, Communication & Society, 15(5), 769–787. DOI:10.2139/ssrn.1926244

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics, 163*(4), 835–850. DOI: 10.1007/s10551-018-3921-3.

Mullane, N. (2019, febrero). La eliminación de los sesgos en los algoritmos. *La Revista de la Normalización Española, 11*.

Noguerol Díaz, Y. (2020). Reseña de la obra «La regulación de los algoritmos» de Alejandro José Huergo Lora (Dir.), Gustavo Manuel Díaz González (Coord.), Aranzadi Thomson Reuters. *Revista Administración & Ciudadanía, 15*(2), pp. 265-266.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), pp. 447-453. DOI: 10.1126/science. aax2342.

Ochigame, R. (2019, diciembre 20). The invention of ethical AI. *The Intercept*. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/ .

Ortiz De Zárate Alcarazo, L. (2022). Explicabilidad (de la inteligencia artificial). *Eunomía. Revista en Cultura de la Legalidad, 22*, p. 338.

Pătraș, L., & Todolí, A. (2022). Ser influencer hoy: Posibilidades y obstáculos de una nueva fuente de empleo. *Papers de la Càtedra d'Economia Colaborativa i Transformació Digital, 4*, p. 54.

Pethig, F., & Kroenung, J. (2023). Biased humans, (un)biased algorithms? *Journal of Business Ethics, 183*(4), 637–652. DOI:10.1007/s10551-022-05071-8.

Pinto Fontanillo, J. A. (2020). *El derecho ante los retos de la inteligencia artificial*. Edisofer.

Savulescu, J. (2012). *¿Decisiones peligrosas? Una bioética desafiante* (Trad. Blanca Rodríguez López & Enrique Bonete Perales). Tecnos.

Savulescu, J., & Maslen, H. (2015). Moral enhancement and artificial intelligence: Moral AI? En J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond Artificial Intelligence: The disappearing human-machine divide* (pp. 79–95). Springer. https://doi.org/10.1007/978-3-319-09668-1_6

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019, Julio). Evaluating gender bias in machine translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1679–1684). Association for Computational Linguistics.

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology, 11*(2), 105–112. DOI:10.1007/s10676-009-9187-9 .

Vantin, S. (2021). Inteligencia artificial y derecho antidiscriminatorio. En F. H. Llano Alonso (Dir.), J. Garrido Martín, & R. D. Valdivia Jiménez (Coords.), *Inteligencia artificial y derecho: El jurista ante los retos de la era digital* (pp. 374-376). Laborum Ediciones.

Vantin, S. (2021). Recensione di Lettieri, N. (2020). *Antigone e gli algoritmi. Appunti per un approccio giusfilosofico.* Mucchi. *Ars Interpretandi, 1*, pp. 195-197.

Referencias Normativas

Grupo de Altos Expertos de la Comisión Europea. (2018). *Ethics guidelines for trustworthy AI*. Comisión Europea. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Parlamento Europeo. (2022). *Artificial intelligence and human rights: Opportunities, risks, and ethical considerations* (EPRS_STU (2022)729541). Servicio de Investigación Parlamentaria Europeo. https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2022)729541

IUS ET SCIENTIA • 2024
Vol. 10 • Nº 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06

Parlamento Europeo. (2024). *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828.* Diario Oficial de la Unión Europea, L XXX/XX.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (2021). *Recomendación sobre la Ética de la Inteligencia Artificial*. https://unesdoc.unesco.org/ark:/48223/pf0000377897.

IUS ET SCIENTIA • 2024
Vol. 10 • N° 2 • pp. 123-151
ISSN 2444-8478 • https://dx.doi.org/10.12795/IESTSCIENTIA.2024.i02.06