InDret
4.2025

Axell Cervera
Universidad de San Martin de Porres
-
Erik Francesc Obiol Anaya
Universidad de San Martin de Porres

Cinco formas de discriminación algorítmica:
Hacia una descripción material

Sumario
-
En los últimos años, la automatización de procesos a través de mecanismos algorítmicos ha experimentado un incremento significativo en sectores como la salud, la educación, el sistema crediticio, el entorno laboral, la vigilancia policial, entre otros. Aunque es evidente que dicha transformación ha traído importantes beneficios en términos de eficiencia y reducción de costos, la implementación de estos sistemas no está exenta de riesgos sociales, en tanto poseen la capacidad de generar dinámicas discriminatorias con un nivel de complejidad mayor a la observada en humanos, lo cual dificulta su comprensión técnica y abordaje analítico. Por este motivo, en el presente trabajo de investigación se examinaron cinco formas en las que puede configurarse la discriminación algorítmica; a saber, discriminación por correlaciones sesgadas, por variables sustitutas, por grupos algorítmicos, por desajuste algorítmico y por bucle de retroalimentación negativo. Posterior a ello, se estableció una propuesta de carácter descriptivo que, por un lado, permite explicar cómo se materializa este fenómeno en la realidad observable, y por el otro, plantea un punto de partida conceptual para futuras investigaciones sobre el tema.

Abstract
-
In recent years, the automation of processes through algorithmic mechanisms has experienced a significant increase in sectors such as healthcare, education, the credit system, the workplace, police surveillance, among others. Although it is evident that this transformation has brought significant benefits in terms of efficiency and cost reduction, the implementation of this systems is not free from social risk, as they have the capacity to generate discriminatory dynamics with a level of complexity greater than that observed in humans, which hinders their technical understanding and analytical approach. For this reason, in the present research work, five ways in which algorithmic discrimination can take shape were examined; namely, discrimination by biased correlations, by proxy variables, by algorithmic groups, by algorithmic mismatch, and by negative feedback loops. Thereafter, a descriptive proposal was established that, on the one hand, allows for explaining how this phenomenon materializes in observable reality, and, on the other hand, provides a conceptual starting point for future research on the subject.

Title: *Five forms of algorithmic discrimination: Towards a material description*
-
Palabras clave: Discriminación algorítmica, Equidad algorítmica, Sesgo algorítmico, Decisiones automatizadas, Derecho antidiscriminatorio, Discriminación por proxy, Bucle de retroalimentación

Keywords**:** *Algorithmic discrimination, Algorithmic fairness, Algorithmic bias, Automated decisions, Anti-discrimination law, Proxy discrimination, Feedback loop*

-
DOI: 10.31009/InDret.2025.i4.17

InDret

4.2025

Recepción
29/07/2025
-
Aceptación
13/10/2025
-

Índice

-

-

1. Introducción*

En 2024, *McKinsey & Company* publicó un informe en el cual se revela que la adopción de la inteligencia artificial en el ámbito corporativo ha experimentado un aumento a nivel global. En comparación con el año 2023, donde los valores no alcanzaban el 66%, la adopción en 2024 se incrementó hasta el 72%.[1] Al igual que en el sector privado, en el ámbito gubernamental las instituciones estatales han comenzado a implementar estos sistemas de última generación con el fin optimizar la eficacia y receptividad de sus servicios.[2] En Reino Unido, por ejemplo, el 22% de 986 profesionales de servicios públicos encuestados indicaron que utilizan activamente un sistema de inteligencia artificial generativa en su trabajo.[3] A principios de esta década, las universidades de *Stanford y New York* realizaron un estudio donde se evaluaron a 142 departamentos, agencias y subagencias federales de Estados Unidos concluyendo que el 45% de estas entidades habían experimentado con sistemas inteligentes.[4] Dicho esto, pese al creciente interés en la materia, la literatura relativa a la adopción de esta tecnología en países emergentes es todavía escasa. No obstante, debido a la globalización es razonable suponer que la tendencia observada en regiones desarrolladas se replicará en dichas naciones durante los próximos años.

Los algoritmos de aprendizaje automático tienen diversas formas de uso en las finanzas, la selección de candidatos para procesos de reclutamiento, la publicidad digital, la atención sanitaria, entre otras varias aplicaciones.[5] Uno de los principales métodos de implementación se da en el proceso de toma de decisiones, esta puede ser laxa, proporcionando recomendaciones, o rígida, a través de decisiones automatizadas.[6] Sobre esto último, el influyente filósofo estadounidense Noam Chomsky, en una entrevista realizada por Mateus Bolson Ruzzarin, sostiene que el riesgo inherente a un proceso incrementa al ser automatizado debido a que estos sistemas pueden fallar al generar falsos positivos.[7] Además, agrega que esto ocurre porque los sistemas inteligentes se centran en el análisis de una enorme cantidad de datos, en la identificación de patrones a partir de ellos y en la utilización de esas relaciones para guiar su funcionamiento. Sin embargo, advierte que tales patrones pueden ser engañosos y que no existe un método que permita automatizar decisiones sensatas.

Un peligro latente de la utilización del aprendizaje automático en el proceso de toma de decisiones, ya sea de manera laxa o rígida, es la discriminación algorítmica. Esta puede definirse, *prima facie*, como el efecto diferencial que un algoritmo impone sobre determinados grupos en

* Autor de contacto: Axell Cervera (axellcervera910@gmail.com)/Erik Francesc Obiol Anaya (eobiola@usmp.pe).
Esta investigación se ha elaborado en el marco académico del "Circulo de Investigación y Argumentación Jurídica" de la Universidad de San Martín de Porres Filial Norte.

[1] SINGLA ET AL., «The state of AI in early 2024: Gen AI adoption spikes and starts to generate value», *QuantumBlack AI by McKinsey*, 2024, pp. 1-22.

[2] STRAUB ET AL., «Artificial intelligence in government: Concepts, standards, and a unified framework», *Computer Science,* 2023, pp. 1-34.

[3] BRIGHT ET AL., «Generative AI is Already Widespread in the Public Sector», 2024, pp. 1-10.

[4] ENGSTROM ET AL., «Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies», 2020, pp. 1-122.

[5] PÁEZ, «Negligent Algorithmic Discrimination», *Law and Contemporary Problems*, 2021, p. 21; RAVISHANKAR ET AL., «Provable Detection of Propagating Sampling Bias in Prediction Models», *The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)*, 2023, pp. 9562-9569.

[6] SCHWARTING/ULBRICHT, «Why Organization Matters in "Algorithmic Discrimination"», *Köln Z Soziol*, 2022, pp. 307-330.

[7] BOLSON RUZZARIN, «Conversation with Noam Chomsky - The Responsibility of Intellectuals», 2021. Véase entre los minutos 30:15-31:07 en: <https://www.youtube.com/watch?v=7M35aasejgI&t=78s>.

función de características protegidas por la ley como el sexo, raza, religión, etc. Sin embargo, como señala Daniel Innerarity, este fenómeno puede ser un tanto abstracto y poco intuitivo en comparación con las formas de discriminación tradicional.[8] Además, la doctrina actual parece carecer de las herramientas necesarias para realizar un análisis comparativo adecuado ya que los algoritmos pueden emplear características aparentemente triviales o excesivamente complejas como base para discriminar.[9]

Por lo mencionado, el propósito de esta investigación es analizar y describir la manera en que los modelos algorítmicos pueden generar, amplificar y perpetuar efectos discriminatorios en la realidad observable. En virtud de ello, la sección 2 ofrecerá una revisión sintética de la literatura académica pertinente, identificando cinco posibles formas de manifestación del fenómeno adverso abordado; a saber, la discriminación por correlaciones sesgadas, por variables sustitutas, por grupos algorítmicos, por desajuste algorítmico y por bucle de retroalimentación negativo. Posteriormente, la sección 3 formulará una propuesta descriptiva de carácter material sobre la discriminación algorítmica, siendo lo *material* una referencia al modo fáctico en que los sistemas afectan negativamente el principio de igualdad y no discriminación.

2. Formas de discriminación algorítmica

En los últimos años, la discriminación algorítmica ha adquirido una notable relevancia en la literatura académica, en tanto representa un problema complejo de naturaleza social, técnica y jurídica que desafía el marco doctrinal establecido sobre las formas tradicionales de discriminación. Si bien la discusión actual se ha centrado en los sistemas inteligentes basados en aprendizaje automático, lo cierto es que el análisis del impacto adverso en cuestión no debe restringirse solo a estos, sino que debe incluir también a toda clase de algoritmos, inteligentes o no, que sean implementados en tareas predictivas o en la toma de decisiones automatizadas. En un sentido elemental, la manera en que los humanos ejecutan dinámicas de exclusión difiere sustancialmente de cómo los sistemas lo hacen, principalmente por la capacidad de estos últimos para detectar correlaciones en los datos, lo que vuelve opaca la cadena causal que condujo al resultado excluyente.

Por este motivo, en las siguientes subsecciones se abordarán cinco supuestos que permiten explicar cómo los modelos algorítmicos son susceptibles de configurar impactos negativos de carácter discriminatorio. Considerando que futuras investigaciones pueden categorizar de manera diferente este problema, en el presente trabajo no se pretende establecer una clasificación taxativa sobre el tema bajo análisis, sino sintetizar los diferentes enfoques teóricos y empíricos existentes en la literatura académica. Se propone que la discriminación algorítmica puede materializarse de las siguientes formas.

2.1. Discriminación por correlaciones sesgadas

El debate académico sobre las implicancias éticas derivadas del uso de tecnologías predictivas en la toma de decisiones suele tomar como referencia inicial el caso *Loomis v. Wisconsin*. En dicho juicio, el tribunal supremo de *Wisconsin* sustentó su sentencia condenatoria utilizando, junto con

[8] INNERARITY, «Justicia Algorítmica y Autodeterminación Deliberativa», *Isegoría Revista de Filosofía moral y política*, núm. 68, 2023, p. 2.
[9] CIRCIUMARU, «Futureproofing EU Law. The Case of Algorithmic Discrimination», *University of Oxford,* 2021, p. 96.

otras herramientas, un informe de evaluación de riesgo generado mediante un algoritmo, el cual indicaba que Loomis presentaba una elevada probabilidad de reincidencia.[10] El modelo en cuestión fue el *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS), desarrollado por la empresa conocida entonces como *Northpointe,* el cual, al analizar 137 parámetros, asignaba una puntuación de 1 a 10 para indicar el nivel de riesgo de reincidencia de los individuos evaluados.[11] En 2016, sin embargo, *ProPublica* demostró que COMPAS reproducía sesgos raciales asignando el doble de probabilidades de ser calificadas como de alto riesgo a personas afrodescendientes.[12] Si bien el caso Loomis no constituye en sentido estricto un supuesto de discriminación algorítmica, este ha permitido visibilizar el riesgo que representa la generación de resultados sesgados por parte de tales sistemas.

De hecho, diversas situaciones donde modelos algorítmicos reproducen resultados basados en prejuicios han sido documentadas: Ziggi Tyler con *TikTok*,[13] Jacky Aluciné con *Google Photos*,[14] traducción sesgada en *Google Translate*,[15] sesgo de género en anuncios de empleo en *Facebook*,[16] entre otras. En todos estos casos, las herramientas automatizadas han sido objeto de crítica por su tendencia a amplificar sesgos sociales preexistentes. Dicho esto, aunque diversos autores utilizan el término *sesgo algorítmico* para describir tal fenómeno, los algoritmos, en esencia, solo reproducen correlaciones sesgadas contenidas en los datos. Como señalan Farič y Bratko, los datos son obtenidos del mundo real donde existen prácticas sesgadas que, mediante principios matemáticos y estadísticos, estos sistemas utilizan para descubrir leyes sobre las cuales regir su funcionamiento, lo cual no significa que los sistemas tengan intenciones maliciosas *per se*.[17]

Mehrabi et al., en su trabajo *«A Survey on Bias and Fairness in Machine Learning»*, identifican y analizan un total de 19 fuentes de sesgos que han sido materia de investigación en la literatura académica.[18] Además, los autores las clasifican en tres categorías que se relacionan de manera cíclica, estas son: 1) Los sesgos trasmitidos de los datos al algoritmo en la fase de entrenamiento, 2) Los trasmitidos del algoritmo al usuario por la modulación del comportamiento de este último al interactuar con los resultados sesgados del primero y 3) Los trasmitidos del usuario a los datos, cuando los sesgos inherentes del primero se ven reflejados en los últimos. Si bien el propósito de este artículo no es abordar de manera exhaustiva cada fuente de manera grupal o individual, es menester utilizar algunas de ellas para ejemplificar cómo los sistemas pueden perpetuar disparidades en los contextos donde son implementados.

[10] ROA AVELLA ET AL., «Uso del Algoritmo COMPAS en el Proceso Penal y los Riesgos a los Derechos Humanos», *RevIsta Brasileira de Direito Processual Penal*, 2022, pp. 275-310.
[11] FARIČ/BRATKO, «Machine Bias: A Survey of Issues», *Informatica*, núm. 48, 2024, pp. 205–212.
[12] SCHWARTING/ULBRICHT, *op. cit.*, p. 317.
[13] GHAFFARY, «How TikTok's Hate Speech Detection Tool Set Off a Debate About Racial Bias on the App», *Vox,* 2021. Obtenido de: <https://www.vox.com/recode/2021/7/7/22566017/tiktok-black-creators-ziggi-tyler-debate-about-black-lives-matter-racial-bias-social-media>, fecha de consulta 10 de marzo de 2025.
[14] GRANT/HILL, «Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's», *The New York Times,* 2023. Obtenido de: <https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html>, fecha de consulta 24 de marzo de 2025.
[15] SÓLMUNDSDÓTTIR ET AL., «Mean Machine Translations: On Gender Bias in Icelandic Machine Translations», *Proceedings of the 13th Conference on Language Resources and Evaluation*, 2022, pp. 3113-3121.
[16] IMANA ET AL., «Auditing for Discrimination in Algorithms Delivering Job Ads», *Proceedings of the Web Conference 2021 (WWW '21)*, 2021, pp. 3767-3778.
[17] FARIČ/BRATKO, *op. cit.*, p. 208.
[18] MEHRABI ET AL., «A Survey on Bias and Fairness in Machine Learning», *ACM Computing Surveys (CSUR)*, 2022, pp. 1-35.

El *sesgo de representación*, ocurre cuando ciertas poblaciones están subrepresentadas en los datos de entrenamiento.[19] Dicho de otro modo, el algoritmo incurre en el citado sesgo cuando, en el proceso de recopilación de datos, la distribución representativa de características sensibles; como la raza, sexo, nacionalidad, condición económica, etc.; es insuficiente en términos de proporción, lo que produce resultados desiguales para individuos pertenecientes a estos grupos. Por ejemplo, en el incidente de Jacky Aluciné, donde *Google Photos* etiquetó a él y a su amigo como gorilas, dos exempleados de *Google* afirmaron que el algoritmo había fallado porque la empresa no incluyó suficientes fotos de personas afrodescendientes en el entrenamiento del modelo.[20] Ahora bien, la subrepresentación puede implicar un riesgo de mayor complejidad cuando se la aborda desde el prisma doctrinal de la *discriminación interseccional*, la cual sostiene que la yuxtaposición de dos o más características sensibles tiende a generar patrones particulares de exclusión que no pueden explicarse mediante el análisis individual de cada factor.[21] En este sentido, la subrepresentación interseccional se configuraría, por ejemplo, si un algoritmo de asignación crediticia no ha sido entrenado con suficientes datos de mujeres afrodescendientes de bajos ingresos económicos o de hombres extranjeros que son testigos de Jehová.

Entendido lo anterior, Salem Alelyani señala que los conjuntos de datos contienen sesgo porque reflejan el comportamiento, la práctica, la experiencia y las acciones humanas.[22] Esto implica que los sesgos son heredados de desigualdades históricas y sistémicas presentes en la realidad, lo cual invita a la reflexión de cómo la subrepresentación es producto, en parte, de dilemas sociales existentes. Si bien el sesgo de representación puede ser inducido por el programador al no entrenar el modelo de manera adecuada, es importante destacar que la complejidad de los algoritmos hace inviable que este tenga en consideración todas las configuraciones sistémicas necesarias para evitar los *bugs* que producen resultados injustos.[23] Esto último, es evidente cuando se reflexiona sobre la subrepresentación interseccional, donde las posibles conjunciones de características son tan diversas que es fáctica y económicamente inviable tenerlas en cuenta. Además, los métodos de balanceo de datos, sobremuestreo de la clase minoritaria y submuestreo de la clase mayoritaria, pueden mejorar el rendimiento de los modelos, pero perjudicar la equidad.[24] Esto se debe, en parte, a que los algoritmos pueden utilizar variables sustitutas de grupos minoritarios, dificultando la eficacia de estos métodos. Estas variables serán abordadas con mayor profundidad en la sección 2.2.

Otro problema existente en algunos algoritmos es el *sesgo de agregación* que, como señalan Mehrabi et al., ocurre cuando estos producen conclusiones erróneas sobre individuos a partir de los patrones observados en toda la población.[25] Para los autores, este tipo de sesgo puede darse, por ejemplo, si una herramienta de apoyo clínico que mide los niveles de HbA1c en sangre para diagnosticar la diabetes no toma en cuenta que el rendimiento de dicho indicador puede variar

[19] SHAHBAZI ET AL., «Representation Bias in Data: A Survey on Identification and Resolution Techniques», *Woodstock '18: ACM Symposium on Neural Gaze Detection,* 2021, pp. 1-47.
[20] GRANT/HILL, *op. cit.*
[21] GARBELLINI FILHO, «El enfrentamiento a la discriminación interseccional en el sistema interamericano de derechos humano. Análisis de las aportaciones del marco de la OEA a la construcción del derecho discriminatorio», en FILLOL MAZO (coord.)*, Los logros de la gobernabilidad en América Latina*, Dykinson S.L., 2024, pp. 127-148.
[22] ALELYANI, «Detection and Evaluation of Machine Learning Bias», *Applied Sciences*, 2021, pp. 1-17.
[23] INNERARITY, *op. cit.*, p. 3.
[24] CHAKRABORTY ET AL., «Bias in Machine Learning Software: Why? How? What to Do?», Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '21), 2021, pp. 429-440.
[25] MEHRABI ET AL., *op. cit.*, p. 5.

dependiendo de la raza.[26] Al replicarse este tipo de sesgo, se podrían perpetuar resultados estigmatizantes en base a características sensibles, donde ciertos individuos, por el solo hecho de pertenecer a una determinada población, son excluidos del acceso a oportunidades laborales, créditos bancarios, etc.

Como se ha señalado, en este articulo no se profundiza de manera exhaustiva todos los posibles sesgos que pueden surgir, sin embargo, resulta importante señalar que las formas de discriminación algorítmica que se desarrollaran en las siguientes secciones derivan en gran parte de las correlaciones sesgadas. Daniel Innerarity resalta que, a pesar de la extendida intención de erradicar toda clase de sesgos, los modelos algorítmicos no pueden operar de manera efectiva sin ellos.[27] El avance tecnológico en las herramientas de aprendizaje automático, los grandes modelos de lenguaje y otros sistemas inteligentes son producto de la evolución en la recopilación, almacenamiento y procesamiento de una inmensa cantidad de datos que están plagados de sesgos. A pesar de que existen estudios que aplican con éxito métodos de mitigación, como el de Han et al.,[28] otras investigaciones, como la realizada por Chen et al.,[29] advierten que los valores de métrica de rendimiento del algoritmo pueden variar dependiendo del método aplicado, incluso, en algunos escenarios, tales técnicas pueden producir valores negativos en términos de equidad en lugar de mejorarla. En cualquier caso, todos estos esfuerzos solo logran mitigar en parte los sesgos y no erradicarlos por completo.

2.2. Discriminación por *proxy* o sustitución

Una propuesta recurrente, cuando se aborda el debate sobre la prevención de la discriminación algorítmica, es la supresión o anonimización de características de protección legal, o sensibles, en los datos de entrenamiento a fin de que tales atributos no influyan en la toma de decisiones del modelo. Sin embargo, esta idea presenta limitaciones, ya que los sistemas de aprendizaje automático correlacionan los atributos protegidos con otras variables no sujetas a restricción, reconstruyendo de manera eficaz la información suprimida.[30] Estas variables, denominadas *proxy* o *variables sustitutas*, contribuyen en la configuración de un fenómeno discriminatorio cuando el algoritmo las emplea para sus predicciones, utilizando características aparentemente neutrales que, debido a su capacidad para sustituir un atributo protegido, pueden generar resultados desfavorables para los miembros de ese grupo. Por ejemplo, como señalan Weerts et al., en ciudades donde los barrios están segregados por líneas étnicas, el código postal puede funcionar como *proxy* de la etnia.[31] Si en dicho entorno un algoritmo utiliza como base el código postal para tomar decisiones sobre la asignación de crédito, es posible que algunos grupos étnicos marginados históricamente se vean afectados de manera negativa.

---

[26] FORD ET AL., «Racial Differences in Performance of HbA1c for the Classification of Diabetes and Prediabetes among US Adults of Non-Hispanic Black and White Race», *U.S. Department of Health and Human Services*, 2019, pp. 1234-1242.

[27] INNERARITY, *op. cit.*, p. 3.

[28] HAN ET AL., «Balancing out Bias: Achieving Fairness Through Balanced Training», 2022, pp. 1-16. En este estudio se aplica un entrenamiento de datos balanceado con enfoque en *Equal Opportunity*, logrando reducir la brecha de verdaderos positivos *(TPR gap)* entre grupos sin sacrificar la precisión competitiva del modelo.

[29] CHEN ET AL., «A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers», *ACM Transactions on Software Engineering and Methodology*, 2023, pp. 1-31. El artículo analiza 17 técnicas de mitigación de sesgos y concluye que, en una fracción significativa de los casos, algunas técnicas reducen el rendimiento y, en determinados contextos, también la equidad.

[30] FARIČ/BRATKO, *op. cit.*, p. 209.

[31] WEERTS ET AL., «Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms», *2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 2024, pp. 1-11.

Cabe resaltar que la discriminación por *proxy* no constituye una manifestación reciente, de hecho, esta práctica ha sido perpetuada de manera intencional por actores humanos a través de la historia. Tal como ilustran Prince y Schwarcz, durante la mitad del siglo XX, los bancos y compañías de seguros utilizaban el *redlining* para eludir diversas leyes antidiscriminatorias relacionadas con la raza.[32] Estas entidades financieras, de manera análoga al ejemplo descrito anteriormente, se abstenían de prestar sus servicios en determinadas zonas geográficas donde existiera una alta concentración poblacional afrodescendiente, utilizando tales territorios como sustitutos de la raza.

En términos generales, es posible afirmar que el desarrollo y aplicación de los algoritmos ha conllevado que el riesgo de discriminación por *proxy* se vea amplificado significativamente. Por un lado, los seres humanos pueden manipular deliberadamente al modelo para que emplee estos sustitutos con el fin de eludir las obligaciones normativas antidiscriminatorias, enmascarando sus conductas perniciosas en la negación de la utilización o conocimiento de dichas variables. Por otro lado, los algoritmos pueden generar resultados discriminatorios empleando *proxies* inadvertidos para los operadores humanos, incluso si se llegara a detectar dicha circunstancia, las correlaciones utilizadas por los sistemas presentan tal complejidad que el establecimiento de mecanismos eficaces de corrección resulta una tarea incierta. En un sentido elemental, y tomando como referencia la perspectiva propuesta por Adams-Prassl et al.,[33] esta modalidad de discriminación se materializa como resultado de decisiones automatizadas basadas en tres categorías de variables *proxy*; estas son, entrenado, aprendido y de variable latente.

Primero, con respecto al *proxy entrenado*, este se hereda directamente de las correlaciones contenidas en los datos introducidos en la fase de desarrollo del modelo. En este sentido, desigualdades estructurales de la sociedad pueden quedar encubiertas en variables *proxy* que, al ser incorporadas en los datos mediante muestras poblacionales extraídas del mundo real, tienen el potencial de inducir resultados negativos en perjuicio de grupos protegidos. Adams-Prassl et al. ilustran claramente este aspecto mediante el ejemplo de una entidad bancaria que implementa un algoritmo para automatizar la evaluación de solicitudes hipotecarias.[34] Si el sistema asocia el matrimonio como un indicador positivo para la asignación del crédito y dicha variable se encuentra incorporada en los datos considerando únicamente uniones heterosexuales, ya sea por razones políticas o por disposiciones restrictivas en la definición de dicha institución jurídica, ello podría dar lugar a un efecto discriminatorio en perjuicio de parejas del mismo sexo, excluidas de tal categoría por impedimentos normativos.

Segundo, el *proxy aprendido*, es aquel que el modelo construye de manera autónoma a partir de la identificación de asociaciones presentes en los datos como producto del análisis computacional de estos. Un caso representativo de este tipo de *proxy* es el del algoritmo de contratación desarrollado, y posteriormente descartado, por Amazon. Dicho sistema otorgaba una calificación de entre una a cinco estrellas a los candidatos, tomando como base el análisis de perfiles curriculares de empleados contratados por la compañía en los diez años previos. Sin embargo, en un reportaje de la prestigiosa agencia de noticias *Reuters*, Dastin reveló que el algoritmo había aprendido de manera autónoma a preferir candidatos masculinos y penalizar los currículos que tuvieran alguna referencia femenina.[35] De manera aún más preocupante, el algoritmo marginaba los perfiles que hubieran egresados de dos universidades que admitían exclusivamente a estudiantes femeninas, utilizando el nombre de dichas instituciones como

[32] PRINCE/SCHWARCZ, «Proxy Discrimination in the Age of Artificial Intelligence and Big Data», *Iowa Law Review*, 2020, pp. 1257-1318.
[33] ADAMS-PRASSL ET AL., «Directly Discriminatory Algorithms», *The Modern Law Review*, 2023, pp. 144-175.
[34] *Ibidem*, P. 158.
[35] DASTIN, «Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women», *Reuters*, 2018. Obtenido de: <https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/>, fecha de consulta 25 de marzo de 2025.

sustituto del género. A pesar de que posteriormente los programadores eliminaron los términos específicos que el sistema utilizó para discriminar, la neutralidad total no estaba garantizada.[36]

Tercero, para explicar cómo los algoritmos pueden discriminar utilizando *proxies de variable latente* se debe empezar comprendiendo el concepto de *variable latente*. Dicho término, hace referencia a los parámetros o atributos inherentemente inobservables que requieren ser inferidos a partir de otros indicadores observables.[37] En el campo de la psicometría, por ejemplo, constructos como la inteligencia, el autocontrol o la agresividad son considerados variables latentes puesto que no pueden inferirse de manera directa, sino que deben estimarse a partir de variables manifiestas como el rendimiento en exámenes, el tiempo en espera tolerado para la recepción de una gratificación o la frecuencia de comportamientos hostiles ante situaciones de estrés. En el caso de los algoritmos, en especial los de aprendizaje automático, la manera en la que las variables latentes influyen en los resultados puede ser de difícil comprensión debido a la gran capacidad de análisis que poseen. Utilizando como muestra un algoritmo de detección de reincidencia, Davies y Douglas sugieren que, como producto del gran procesamiento computacional del modelo, este podría utilizar una amplia combinación de factores para capturar la correlación entre la raza y la reincidencia, de modo que, aunque el parámetro raza hubiere sido excluido de los datos, su relación con la reincidencia seguirá funcionando como una variable latente.[38] Es decir, a pesar de que el sistema no haya utilizado directamente alguna característica protegida como factor determinante para su resultado, este sí puede haberla reconstruido de manera indirecta como una variable latente y utilizado características inentendibles como proxies perfectos de esta. En otras palabras, el algoritmo estaría generando discriminación a partir de *proxies* de un atributo protegido que ha sido excluido como parámetro causal del resultado, pero que sigue funcionando como una variable latente influyente en la decisión.

Siguiendo con la categoría racial como caso ilustrativo, Lily Hu sostiene, a la luz de lo expuesto con anterioridad, que un sistema algorítmico puede discriminar en función de la raza tanto si la incluye explícitamente en su proceso de decisión como si, prescindiendo de ella, opera mediante factores constitutivos de la misma para llegar a un resultado.[39] En virtud de ello, la autora propone adoptar un enfoque *constructivista grueso*, en el sentido de no apreciar el atributo protegido solamente como una mera etiqueta, sino entenderlo como un conjunto de relaciones estructurales del mundo real que lo aproximan. Dado la complejidad que representa la discriminación algorítmica por *proxy*, posturas como esta parecen brindar herramientas conceptuales que pueden ser utilizadas para el desarrollo de mecanismos de mitigación contra los impactos discriminatorios bajo análisis. No obstante, resultar complejo establecer con certeza los factores que han sido aprendidos por el modelo como perfectamente constitutivos de un atributo sensible en base al cual discrimina, especialmente en aquellos sistemas conocidos como de *caja negra (black box)*, que como señala Pavlidis, operan con un alto grado de opacidad.[40] Además, debido al poder predictivo no medible de la inteligencia artificial, Prince y Schwarcz, advierten que la discriminación por *proxy* es prácticamente inevitable.[41]

---

[36] IRIONDO, «Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women», *Carnegie Mellon University*, 2018. Obtenido de: <https://ml.cmu.edu/news/news-archive/2018/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women>, fecha de consulta 25 de marzo de 2025.

[37] BORGSTEDE/EGGERT, «Squaring the Circle: From Latent Variables to Theory-Based Measurement», *Theory & Psychology*, 2023, pp. 118-137.

[38] DAVIES/DOUGLAS, «Learning to Discriminate: The Perfect Proxy Problem in Artificially Intelligent Sentencing», *Sentencing and Artificial Intelligence*, 2022, pp. 97-121.

[39] HU, «What is "Race" in Algorithmic Discrimination on the Basis of Race?», *Journal of Moral Philosophy*, 2023, pp. 1-23.

[40] PAVLIDIS, «Unlocking the Black Box: Analysing the EU Artificial Intelligence Act's Framework for Explainability in AI», *Law Innovation and Technology*, vol. 16, núm. 1, 2024, pp. 293-308.

[41] PRINCE/SCHWARCZ, *op. cit.*, p. 13.

2.3. Discriminación por grupos algoritmicos

Hasta ahora, los dos casos previos han sido desarrollados en función del perjuicio discriminatorio ejercido contra grupos tradicionalmente protegidos por las normas legales. Como se ha indicado de manera previa, las también denominadas características sensibles comprenden atributos tales como la raza, el sexo, la religión, la orientación sexual o la discapacidad. Desde un plano jurídico internacional, dichos grupos suelen estar representados en listas enunciativas en constituciones, tratados, reglamentos, leyes, entre otros, que sirven para identificar conductas sospechosas de discriminación.[42] Sin embargo, los algoritmos, en especial los de aprendizaje automático, pueden discriminar a partir de grupos mucho más sofisticados sin recaer en los que son protegidos por la ley o en variables *proxies* de ellos. Dicho de otro modo, pueden generarse nuevos grupos vulnerables a partir de clasificaciones algorítmicas que no encajan dentro de los límites del derecho antidiscriminatorio.[43]

Como señala Wahid, en el contexto de las redes sociales los sistemas suelen estar diseñados: *en base al contenido*, utilizando el perfil del usuario para vincularlo con contenido acorde a su interés; *conscientes del contexto*, ajustando sus resultados extrayendo datos personales como la ubicación; y *empleando el aprendizaje automático*, modulando sus recomendaciones de manera continua.[44] Las empresas propietarias programan dichos modelos con el fin de maximizar la participación del usuario en la exposición publicitaria, obteniendo un gran beneficio económico por ello. La segmentación desde este prisma genera efectos positivos en determinados contextos, no obstante, resulta preocupante que este tipo de diseño algorítmico se utilice para la toma de decisiones sobre la asignación crediticia, la contratación de personal, el control fronterizo o el diagnóstico médico. De hecho, en muchas ocasiones la segmentación puede resultar tan absurda que no justifica que cierto grupo de personas, que no encajan dentro del marco regulatorio tradicional, sean privadas del acceso equitativo a recursos y oportunidades.

Un caso representativo del tema en cuestión es la plataforma de contratación impulsada por inteligencia artificial de *Hirevue*. Dicha empresa está especializada en el apoyo técnico de recursos humanos relacionado con la selección de personal mediante el análisis computacional de patrones de habla y expresiones faciales de los candidatos entrevistados en formato de video.[45] En 2019, el *Electronic Privacy Information Center* interpuso una queja formal ante la Comisión Federal de Comercio de los Estados Unidos argumentando que *Hirevue* basaba el 29% de la puntuación de los candidatos en los movimientos faciales, lo que consideraban constituía un riesgo considerable de producción de resultados adversos.[46] La gesticulación del rostro o el movimiento ocular no son atributos que la legislación actual haya previsto como causales de discriminación, sin embargo, el caso bajo análisis abre una brecha de reflexión importante sobre cómo los modelos pueden interpretar tales rasgos como factores influyentes en la decisión final.

Durante una conferencia realizada por el *Weizenbaum Institute*, la profesora Sandra Wachter señaló que, en relación con la toma de decisiones, los algoritmos inteligentes tienen la capacidad de generar agrupaciones a partir de criterios diferentes a los tradicionales, creándolas a partir de

[42] SALOMÉ RESURRECCIÓN, «La discriminación y algunos de sus calificativos: directa, indirecta, por indiferenciación, interseccional (o múltiple) y estructural», *Pensamiento Constitucional*, 2017, p. 261.
[43] TEO, «Artificial intelligence and its 'slow violence' to human rights», *AI and Ethics*, 2024, p. 12.
[44] WAHID, «How to Get Away With Discrimination: The Use of Algorithms to Discriminate in the Internet Entertainment Industry», *American University Journal of Gender, Social Policy & the Law*, 2023, pp. 107-139.
[45] HIREVUE, «About Us», *Hirevue*, 2025. Obtenido de: <https://www.hirevue.com/about>, fecha de consulta 26 de marzo de 2025; RAMÍREZ-BUSTAMANTE/PÁEZ, «Análisis Jurídico de la Discriminación Algorítmica en los Procesos de Selección Laboral», 2021, p. 3.
[46] PÁEZ, *op. cit.*, p. 22-23.

variables como el movimiento del *mouse*, el uso de mayúsculas en la redacción, la velocidad cardiaca o la frecuencia en el parpadeo.[47] El sistema de crédito social chino, por ejemplo, califica como indicadores negativos en la puntuación asignada conductas como el jugar videojuegos en exceso, la realización de gastos innecesarios o el fumar en zonas libres de humo.[48] Aunque dicho sistema no ha sido implementado de manera general, es controversial que en ciertos sectores se utilicen estos parámetros para etiquetar a personas como *individuos desacreditados* dentro de una lista negra, restringiéndoles el uso de vuelos y trenes de alta velocidad.[49]

En este sentido, las agrupaciones no convencionales utilizadas por los algoritmos para privar el acceso equitativo a recursos y oportunidades son denominadas por Wachter como *grupos algorítmicos*, los cuales clasifica en: 1) Los *comprensibles* para el ser humano, pero que no gozan de protección bajo las leyes antidiscriminatorias, como los dueños de gatos, aficionados al fútbol o cinéfilos; y, 2) Los *incomprensibles,* que no pueden ser interpretados por los humanos debido a su gran complejidad, como el historial de la web, la velocidad de *scrolling* o los píxeles individuales de una imagen.[50] En esencia, aunque los grupos algorítmicos no encajan dentro de la lista enunciativa que la ley designa como grupos protegidos, el efecto algorítmico perjudicial generado sobre ellos es funcionalmente equiparable. En primer lugar, porque resulta absurdo que decisiones que repercuten en el proyecto de vida de las personas estén supeditada a factores como los descritos en la clasificación anterior. Y, en segundo lugar, porque dichas agrupaciones no se encuentran dentro del control de las personas afectadas, pues no tuvieron un papel activo en su inclusión dentro de ellas y no tienen la potestad de poder cambiarse a otra o salirse de ellas a voluntad.

Un sector de la doctrina sostiene que los atributos protegidos lo son en tanto representan un carácter inmutable para quienes los poseen, haciendo que sea moralmente cuestionable que se genere una desventaja en perjuicio de su portador por el solo hecho de importarlos. Para Barocas, el uso de atributos inmutables como factor base en la toma de decisiones constituye una amenaza para la autonomía personal debido a que escapan de la elección o control del individuo.[51] Además, si bien algunas características pueden ser aparentemente modificables, como la condición económica o lugar de residencia, en la realidad tal modificación está limitada por situaciones financieras, culturales o sociales que las convierte en inmutables de manera funcional. En contraposición, autores como Foran sostienen que la protección legal a ciertos grupos parte de la irrelevancia moral y la saliencia social, advirtiendo que es conceptualmente limitante interpretar la inmutabilidad como el único criterio susceptible de protección.[52] Sin embargo, aunque la irrelevancia moral es un criterio necesario en el análisis de los grupos algorítmicos, resulta problemático afirmar que estos son socialmente salientes en la interacción entre individuos, sobre todo aquellos que representan una alta complejidad de comprensión para los seres humanos. Por dicho motivo, se puede argumentar que los grupos algorítmicos merecen protección de las leyes antidiscriminatorias debido a que, además de ser moralmente irrelevantes, funcionan como características inmutables producto de la opacidad, ambigüedad, inestabilidad, involuntariedad e inexplicabilidad de los factores que los algoritmos toman como base para generar decisiones respecto al acceso a oportunidades, bienes y servicios. Como

[47] WEIZENBAUM-INSTITUT, «Sandra Wachter - The Theory of Artificial Immutability», 2023. Véase entre los minutos 3:24-4:05 en: <https://www.youtube.com/watch?v=khw1aXupTgk>.
[48] NITTLE, «Spend "Frivolously" and Be Penalized Under China's New Social Credit System», *Vox,* 2018. Obtenido en: <https://www.vox.com/the-goods/2018/11/2/18057450/china-social-credit-score-spend-frivolously-video-games>, fecha de consulta 27 de marzo de 2025.
[49] CHEN/GROSSKLAGS, «Social Control in the Digital Transformation of Society: A Case Study of the Chinese Social Credit System», *Social Sciences*, vol. 11, 2022, pp. 1-23.
[50] WACHTER, «Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law», *Tulane Law Review*, 2022, pp. 5-6.
[51] BAROCAS ET AL., «Fairness and Machine Learning. Limitations and Opportunities», 2023, p. 41.
[52] FORAN, «Grounding Unlawful Discrimination», *Legal Theory*, 2022, pp. 10-11.

sostiene Wachter, tales agrupaciones adquieren una condición de *inmutabilidad artificial* debido a que son producto del cálculo computacional del algoritmo, circunstancia que impone la necesidad de asignarles un tratamiento jurídico equiparable al de características inmutables naturales tuteladas por el derecho antidiscriminatorio.[53]

2.4. Discriminación por desajuste algorítmico

Como se ha desarrollado en las tres subsecciones precedentes, la discriminación algorítmica es un fenómeno que transgrede el derecho fundamental de la igualdad generando resultados injustos en contra de grupos protegidos, variables *proxies* de estos y grupos algorítmicos no protegidos, pero que son moralmente irrelevante e inmutables de manera artificial. Si bien es cierto, el debate en torno al impacto negativo de los algoritmos se ha focalizado principalmente en los sesgos presentes en los datos y las correlaciones espurias aprendidas a partir de estos, no es menos cierto que dicha perspectiva resulta parcial e insuficiente para explicar el fenómeno discriminatorio bajo análisis. Además de lo señalado, los efectos algorítmicos adversos pueden originarse también producto del diseño técnico subyacente del modelo, que al interactuar con el entorno en el cual se implementa, genera resultados no previstos que afectan de manera injusta a ciertos grupos o contextos sociales.

Como señala Wörle, la interacción del algoritmo con el entorno puede configurarse de dos maneras.[54] Por un lado, el modelo puede ser *estático* cuando opera de manera invariable, procesando únicamente los datos de entrenamiento sin ajustarse a los cambios del entorno. Este tipo de configuración resulta útil, por ejemplo, para la asistencia de decisiones clínicas,[55] la gestión de recursos humanos[56] o la detección de fraudes financieros.[57] En contraste, el modelo puede ser *dinámico* cuando se adapta a los potenciales cambios del entorno en el cual se aplica, dicha adaptabilidad se logra utilizando un bucle de retroalimentación donde los resultados observados en la realidad y la información contextual obtenida se reintroducen como nuevos datos. Su implementación es frecuente en sistemas de recomendación,[58] en el ruteo vehicular[59] o en la optimización de portafolios financieros.[60] En la presente subsección se abordarán los impactos discriminatorios provocados por el desajuste entre el algoritmo, implementado de manera estática, y el entorno en el cual opera.

El concepto de *drift* (deriva) es utilizado en la literatura académica para describir los cambios en la distribución subyacente de los datos de un proceso observado, como resultado de las modificaciones en el entorno a lo largo del tiempo.[61] En términos sencillos, el algoritmo puede,

[53] WACHTER, *op. cit.*, p. 48.
[54] WÖRLE, «Negative Feedback Loops and Self-fulfilling Prophecies: Sociotechnical Assessment of Unfairness in Predictive Algorithms», *Department of Computer Science Ludwig-Maximilians-Universität at München*, 2024, pp. 13-18.
[55] PAPADOPOULOS ET AL., «A Systematic Review of Technologies and Standards Used in the Development of Rule Based Clinical Decision Support Systems», *Health and Technology*, 2022, pp. 713-727.
[56] NOWAK, «The Impact of Rule Based Decision Engines on Business Efficiency», *Higson*, 2024. Obtenido de: <https://www.higson.io/blog/the-impact-of-rule-based-decision-engines-on-business-efficiency>, fecha de consulta 28 de marzo de 2025.
[57] MALIK ET AL., «Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture», *Mathematics*, 2022, pp, 1-16.
[58] ALFAIF, «Recommender Systems Applications: Data Sources, Features, and Challenges», *Information*, 2024, pp. 1-25; HASAN ET AL., «Review-based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives», *Proceedings of the ACM Measurement Anaysis of Computing Systems*, 2024, pp. 1-33.
[59] GAMA ET AL., «Multi-Agent Environments for Vehicle Routing Problems», 2024., pp. 1-17.
[60] YUAN ET AL., «Your Offline Policy is Not Trustworthy: Bilevel Reinforcement Learning for Sequential Portfolio Optimization», 2025, pp. 1-21.
[61] HINDER ET AL., «One or Two Things We Know About Concept Drift. A Survey on Monitoring Evolving Environments», 2023, pp. 1-44; HOVAKIMYAN/BRAVO, «Evolving Strategies in Machine Learning: A Systematic Review of Concept Drift Detection», *Information*, 2024, pp. 1-24.

en un primer momento, estar bien calibrado para desempeñar ciertas operaciones en una determinada población objetivo, pero dicha operabilidad se tornará defectuosa en la medida que la distribución en los datos poblacionales cambie como producto del paso del tiempo. Al respecto, Barocas advierte que si un modelo no es reentrenado en función de las transformaciones del contexto en donde es aplicado, existe el riesgo de que se produzcan resultados contrarios a la equidad.[62] Por ejemplo, un algoritmo programado para alertar sobre posibles accidentes de tránsito que basa sus predicciones en la detección de signos de somnolencia en los conductores de una localidad, como ojos semicerrados o movimientos de cabeza. Si, como producto de procesos migratorios, la población objetivo experimenta una mayor presencia de individuos con rasgos fenotípicos faciales característicos de sociedades del Este Asiático, y el modelo no presenta mecanismos de adaptación, se podrían generar falsos positivos discriminatorios en detrimento de este nuevo grupo étnico incorporado.

Otra forma en la que puede surgir un desajuste algorítmico es cuando este, que ha sido diseñado y optimizado para funcionar en un entorno determinado, es aplicado en un contexto diferente de aquel que sirvió como base para su entrenamiento, lo que puede provocar errores significativos en las predicciones de este nuevo entorno. Las predicciones son correlaciones que, mediante un análisis computacional, el sistema utiliza para brindar resultados en un momento y lugar determinado, lo que las hace susceptibles a variaciones contextuales.[63] Por ejemplo, si se emplean datos de justicia penal provenientes de una jurisdicción territorial específica para predecir el riesgo de reincidencia en individuos pertenecientes a otra, es posible que los resultados sobre las tasas de reincidencia para determinados grupos demográficos varíen en función de la localidad.[64]

En campos como el de la psicología, se ha evidenciado que las muestras de datos proceden, en su mayoría, de sociedades occidentales, educadas, industrializadas, ricas y democráticas, conocidas como WEIRD por sus siglas en inglés.[65] Puede que algoritmos entrenados para la autoevaluación o el tamizaje psicológico resulten eficaces para dichas poblaciones, pero generen resultados defectuosos y discriminatorios en grupos demográficos que no encajan dentro de las variables WEIRD. Caso análogo se presentaría si un sistema de toma de decisiones es implementado de la misma forma en países latinoamericanos y en sociedades del norte global, puesto que la percepción sobre constructos como la raza difieren significativamente entre ambas. Este fenómeno, llamado *daltonismo racial* por la doctrina, representa un gran obstáculo para la detección de parámetros o *proxies* que sean indicadores clave de resultados sesgados, dado que la realidad latinoamericana presenta particularidades culturales y estructurales incompatibles con las de sociedades euroatlánticas en las que comúnmente se realiza el desarrollo de los algoritmos.[66]

En términos generales, a menos que el modelo presente algún mecanismo de adaptabilidad en su diseño, este no debe ser aplicado indistintamente en realidades poblacionales heterogéneas, debido al potencial peligro que el desajuste representa para la equidad. Esta reflexión parte de la premisa según la cual el principio de igualdad no solo es transgredido cuando se trata desigualmente condiciones iguales, sino que también cuando se otorga un tratamiento igualitario a circunstancias que, a razón de sus diferencias sustanciales, requieren ser tratadas de manera desigual, lo que ocasiona un daño y ha sido catalogado por la doctrina como

[62] BAROCAS ET AL., *op. cit.*, p. 13.
[63] MUÑOZ GUTIÉRREZ, «La Discriminación en una Sociedad Automatizada: Contribuciones desde América Latina», *Revista Chilena de Derecho y Tecnología*, 2021, p. 281.
[64] RAVISHANKAR ET AL., *op. cit.*, p. 9562.
[65] ANDRINGA/GODFROID, «Sampling Bias and the Problem of Generalizability in Applied Linguistics», *Annual Review of Applied Linguistics*, 2020, p. 134.
[66] MUÑOZ GUTIÉRREZ., *op. cit.*, p. 292.

discriminación por indiferenciación.[67] Aunque a nivel legislativo la incorporación expresa de este argumento ha sido tomada con reticencia en virtud del principio de generalidad de la ley, tanto en la doctrina como en la jurisprudencia se ha establecido que, en determinados supuestos, es necesaria la adopción de este enfoque para no dejar en desprotección a personas o situaciones que merecen un trato distinto. En el contexto de la implementación algorítmica, resulta crucial para el debate académico examinar los riesgos de discriminación asociados al desajuste generado por una *aplicación indiferenciada* del modelo en entornos que son significativamente diferentes.

2.5. Discriminación por bucle de retroalimentación negativo

Como se ha señalado previamente, los algoritmos dinámicos son aquellos que, a diferencia de los implementados de manera estática, tienen la capacidad de adaptarse en el contexto donde operan gracias a mecanismos de retroalimentación o actualización continua. Tales mecanismos utilizan de manera cíclica los resultados generados por el modelo como nuevos datos de entrada que influyen directa o indirectamente en los resultados futuros, dando lugar a lo que diversos investigadores han denominado *bucle de retroalimentación*.[68] Debido a la estructura adaptativa orientada al contexto, la implementación dinámica puede contribuir en la mitigación de los riesgos asociados a la discriminación y la reproducción de sesgos que presentan los algoritmos estáticos, mejorando con ello tanto el rendimiento del modelo como la eficiencia económica de este.[69] Sin embargo, como han advertido Dolata et al., en determinados escenarios la retroalimentación cíclica puede convertirse en una fuente de injusticia introduciendo efectos no deseados que a largo plazo impactan de manera negativa el entorno en donde opera.[70]

Los sistemas de predicción policial han sido objeto de análisis recurrente en relación con los impactos perjudiciales que el bucle bajo análisis puede generar. En diversas investigaciones, como la realizada en 2019 por el Centro de Ética e Innovación de Datos del gobierno del Reino Unido, se ha observado que el etiquetar ciertas zonas como puntos críticos de criminalidad condiciona el comportamiento de los agentes, quienes desarrollan expectativas de conflicto volviéndose más propensos a detener o arrestar personas en dichos territorios en base a prejuicios y no en una necesidad operativa real.[71] Si, *exempli gratia*, un departamento policial utiliza un algoritmo predictivo que categoriza como de alto riesgo ciertos vecindarios donde existe una alta concentración de migrantes, esto podría derivar en un aumento en el despliegue de agentes del orden en tales sectores urbanos. Este incremento de la actividad policial intensificaría la probabilidad arrestos sesgados que, al introducirse como nuevos datos en el sistema, darían paso a nuevos resultados que materializan un *bucle de retroalimentación negativo*

[67] LANDA ARROYO, «El Derecho Fundamental a la Igualdad y No Discriminación en la Jurisprudencia del Tribunal Constitucional del Perú», *Estudios Constitucionales*, vol. 19, núm. 12, 2021, pp. 86-87; MATA SIERRA, «La Discriminación por Indiferenciación y su Incidencia en el Ámbito Tributario», *Revista Jurídica de la Universidad de León*, núm. 8, 2021, pp. 185-201.

[68] PAGAN ET AL., «A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems», 2023, pp. 1-2; TAPIA LARA, «A Conceptual Framework for Simulating Feedback Loops in Engineering Design», *The University of Leeds*, 2023, pp. 45-46.

[69] BAUER ET AL., «Feedback Loops in Machine Learning: A Study on the Interplay of Continuous Updating and Human Discrimination», *Journal of the Association for Information Systems*, vol. 25(4), 2024, p. 818; WÖRLE, *op. cit.*, p. 24.

[70] DOLATA ET AL., «A Sociotechnical View of Algorithmic Fairness», *Information Systems Journal*, 2021, p. 36.

[71] HEAVEN, «Predictive Policing Algorithms Are Racist. They Need to Be Dismantled», *MIT Technology Review*, 2020. Obtenido de: <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/>, fecha de consulta 5 de abril de 2025.

discriminatorio en perjuicio de estos grupos minoritarios, catalogándolos como peligrosos y empezando el ciclo desde el punto inicial.

Las preocupaciones respecto a este fenómeno han alcanzado también el ámbito de la gestión de recursos humanos. En la actualidad, la empresa BMW está desarrollando un algoritmo con retroalimentación destinado a la identificación de candidatos con alto potencial para ocupar roles de liderazgo, seleccionándolos para un programa de capacitación especial. Wörle advierte que los candidatos elegidos, al tener acceso a dicho entrenamiento, incrementan significativamente sus probabilidades de tener éxito y ser promovidos, lo cual refuerza aquellos perfiles que el algoritmo favoreció desde un inicio.[72] Dado que los cargos de liderazgo en la industria automotriz han sido ocupados históricamente por hombres, el ciclo algorítmico podría acentuar en gran medida las desventajas de género. Por otro lado, en el área de la salud, Adam et al. indican que si un modelo predice de manera errónea que un paciente morirá y el personal médico toma como válida dicha predicción, se le serian administrados cuidados paliativos en lugar del tratamiento adecuado, lo que posteriormente conduciría a su fallecimiento.[73] Al actualizarse el modelo con la incorporación este nuevo caso, el sesgo inicial se vería reforzado, desencadenando una mayor tasa de falsos positivos generando resultados realmente alarmantes, sobre todo en contextos críticos como en unidades de cuidados intensivos donde los recursos son limitados y la priorización de pacientes es esencial. Si bien no puede argumentarse que este último ejemplo es un caso de discriminación concreto, su análisis es importante para demostrar las implicancias negativas que generan estos ciclos de retroalimentación.

Los efectos indeseados del bucle no son atribuibles únicamente al funcionamiento de los algoritmos, sino que también a la forma en la que los humanos interpretan y gestionan los resultados. De hecho, la interacción de estos con las predicciones configura, en la práctica, una dinámica cíclica que puede explicarse mediante; la confianza, la empleabilidad y la autojustificación. En primer lugar, en palabras de Cabiddu et al., la confianza inicial que desarrollan los usuarios se debe a su predisposición a confiar en los algoritmos, su diseño que emula el comportamiento humano y la percepción positiva tanto de su utilidad como de su facilidad de uso.[74] Debido a esto, las personas confían en los resultados incluso cuando no deberían hacerlo, aunque sean capaces de hacer la tarea correctamente por sí mismos y hayan percibido indicadores que sugieren la poca fiabilidad de la recomendación.[75] En segundo lugar, como consecuencia de considerarlas lo suficientemente creíbles, los individuos emplean las predicciones, utilizándolas como base decisional para la realización de acciones u operaciones.[76] Aunque dicha práctica puede estar completamente automatizada, lo esencial en este análisis es resaltar que la empleabilidad de las predicciones produce un cambio en el mundo real, introduciendo sesgos y perpetuando efectos discriminatorios. En el caso comentado de predicción policial, el incremento de arrestos hacia los migrantes resulta un claro ejemplo de ello. En tercer lugar, este cambio gradual en el ecosistema convierte en verdad las predicciones,

[72] WÖRLE, *op. cit.*, pp. 54-72.
[73] ADAM ET AL., «Hidden Risks of Machine Learning Applied to Healthcare: Unintended Feedback Loops Between Models and Future Data Causing Model Degradation», *Proceedings of Machine Learning Research*, 2020, p. 2.
[74] CABIDDU ET AL., «Why do Users Trust Algorithms? A Review and Conceptualization of Initial Trust and Trust Over Time», *European Management Journal*, vol. 40, núm. 5, 2022, pp. 685-706.
[75] SURESH ET AL., «Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making», *12th ACM Conference on Web Science*, 2020, pp. 314-324.
[76] KING/MERTENS, «Self Fulfilling Prophecy in Practical and Automated Prediction», *Ethical Theory and Moral Practice*, 2023, pp. 134-135.

las cuales son tomadas como referencia para futuras decisiones del modelo.[77] Este proceso hace que la predicción, aunque errónea en un inicio, se auto justifique, legitimando el funcionamiento del modelo, las futuras recomendaciones que brindara y generando en los usuarios una confianza a largo plazo.

Extendiendo el análisis de este último punto, para Adams-Prassl et al. la autojustificación de los resultados puede generar un preocupante *status quo* discriminatorio.[78] Dado que las predicciones inciden en la realidad, el comportamiento humano se moldea en función de ellas no solo modificando la distribución de los datos que posteriormente serán recopilados por el mecanismo de retroalimentación, sino también incorporando a tales datos sus sesgos y conductas discriminatorias inherentes. En este sentido, la nueva realidad contaminada de injusticia influirá de manera determinante en el modelo, auto justificando la discriminación al aumentar la probabilidad de que esta se convierta en una *profecía autocumplida* (PAC). Conforme a la descripción desarrollada por King y Mertens, una PAC se entiende como aquella predicción cuya utilización contribuye activamente en la concreción del resultado que anticipó.[79] Los casos previamente analizados en materia de predicción policial y gestión de recursos humanos ponen en manifiesto cómo el propio empleo de los resultados coadyuva en la generación de PACs que perpetúan efectos discriminatorios. De este modo, el incremento del patrullaje en zonas habitadas por migrantes refuerza el estigma social que los asocia a la peligrosidad, mientras que la selección exclusiva de perfiles masculinos para programas de capacitación impide que candidatas femeninas desarrollen las habilidades necesarias para ser catalogadas como aptas para roles de liderazgo en el futuro.

Resulta pertinente señalar que las PACs tienen el potencial de enmascaras los errores y alterar la realidad, de modo que no solo se limitan a brindar una predicción sobre el proceso observado, sino que influyen en todos los factores necesarios para que dicha predicción devenga en verosímil. Este fenómeno, resulta análogo a la denominada *tesis de la performatividad* en economía, la cual sostiene que las teorías que intentan explicar los mercados, en la práctica, terminan moldeándolos.[80] Tal carácter performativo de los ciclos algorítmicos dificulta la detección de impactos adversos a la equidad, lo que a su vez obstaculiza la explicabilidad de los resultados, especialmente cuando los modelos utilizan variables *proxy* y grupos algorítmicos inentendibles que carecen de transparencia causal. Sumando a ello, la mitigación de discriminación algorítmica por bucle de retroalimentación requiere, de manera crítica, la ausencia de comportamientos discriminatorios por parte de los tomadores de decisiones humanos; caso contrario, dicho impacto no solo persistirá, sino que se verá amplificado de manera considerable.[81]

3. Hacia una descripción material

Como se ha mencionado en la parte introductoria, el nudo problemático que aquí se aborda puede ser definido, *prima facie*, como el efecto diferencial que un modelo impone sobre determinados grupos en función de características protegidas legalmente. De hecho, esta perspectiva ha sido formalmente reconocida en diversos países como Estados Unidos, donde la Casa Blanca *(The*

[77] WYLLIE ET AL., «Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias», 2024, p. 2.
[78] ADAMS-PRASSL ET AL., *op. cit.*, pp. 153-154.
[79] KING/MERTENS, *op. cit.*, p. 128.
[80] WYLLIE ET AL., *op. cit.*, p. 1.
[81] BAUER ET AL., «Feedback Loops in Machine Learning: A Study on the Interplay of Continuous Updating and Human Discrimination», *Journal of the Association for Information Systems*, vol. 25(4), 2024, P. 819.

White House) ha definido, en *«The Blueprint for an AI Bill of Rights»,* que la discriminación algorítmica «ocurre cuando los sistemas automatizados contribuyen a un trato diferente injustificado o a impactos que desfavorecen a personas en función de su raza, color, etnia, sexo, religión, edad, origen nacional, discapacidad, condición de veterano, información genética o cualquier otra clasificación protegida por la ley».[82] Si bien la mencionada noción carece de obligatoriedad federal, su influencia regulatoria se ha manifestado en el plano legislativo de *Colorado*, donde ha sido incluida en la sección 6-1-1701(1)(a) del SB 24-205, vigente como ley estatal desde el 2026.[83]

En esencia, el enunciado establecido por la Casa Blanca resulta acertado en términos generales, sin embargo, su composición no contempla el hecho de que los sistemas automatizados tienen la capacidad de discriminar utilizando variables *proxy* o generando agrupaciones complejas. Dicha omisión es particularmente notable, considerando que el propio *«Blueprint for an AI Bill of Rights»* ha dedicado un apartado a subrayar la necesidad de una protección proactiva ante los *proxies*, sugiriendo además que su utilización en el proceso de toma de decisiones podría constituir una práctica legalmente prohibida. Bajo este argumento, el Fiscal General y la División de Derechos Civiles de *New Jersey* han señalado, sin ofrecer un marco conceptual que integre explícitamente tales variables, que el uso de *proxies* puede constituir un supuesto de trato diferencial al aplicar su ley antidiscriminatoria estatal en casos de discriminación algorítmica,[84] lo cual destaca la necesidad de establecer un concepto integral del fenómeno estudiado.

La Unión Europea, en su Reglamento de Inteligencia Artificial (RIA),[85] ha contemplado algunos de los riesgos discriminatorios desarrollados en la presente investigación: sesgos, en los considerandos 67 y 110, así como en los artículos 10.2 y 10.5; desajuste algorítmico, en el considerando 31; y bucle de retroalimentación negativo, en el considerando 67 y el artículo 15.4. En un sentido elemental, la RIA ha supuesto un avance significativo en la regulación de los modelos algorítmicos, en particular de aquellos basados en aprendizaje automático; no obstante, el abordaje de la discriminación algorítmica se encuentra diseminado en el cuerpo normativo y no se incorpora una definición unívoca de este problema. Ello podría, en determinados escenarios, acarrear el surgimiento de discrepancias interpretativas dentro de la Unión sobre qué circunstancias, y cuales no, encajarían dentro de un supuesto de discriminación generada por algoritmos, teniendo cada país que apoyarse en su respectiva normativa nacional.

En España, por ejemplo, tanto la Ley 15/2022, de 12 de julio, (art. 26)[86] como la Carta de Derechos Digitales (apartado VIII)[87] se limitan a abordar el problema procurando la minimización o

[82] El *«The Blueprint for an AI Bill of Rights»* es un documento oficial emitido por el *White House Office of Science and Technology Policy* en octubre del 2022. En él se establecen una serie de principios que buscan proteger al público estadounidense frente a los avances de la inteligencia artificial. El texto disponible para su descarga en: <https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>.

[83] El SB 24-205 es un proyecto de ley aprobado por la *Colorado General Assambly* y firmado en mayo del 2024 por el Gobernador Jared Polis. Su entrada en vigor se postergó hasta el año 2026 con el fin de que las empresas y el propio aparato legislativo tengan tiempo para construir un marco técnico y jurídico idóneo para su implementación. Disponible en: <https://leg.colorado.gov/sites/default/files/2024a_205_signed.pdf>.

[84] PLATKIN/IYER, «Guidance on Algorithmic Discrimination and the New Jersey Law Against Discrimination», 2025, p. 10. Guía emitida por el Fiscal General de *New Jersey* y la División de Derechos Civiles con el propósito de aclarar cómo la ley antidiscriminatoria estatal *(New Jersey Law Against Discrimination)* es aplicable a la discriminación algorítmica. Texto disponible para su descarga en: <https://www.nj.gov/oag/newsreleases25/2025-0108_DCR-Guidance-on-Algorithmic-Discrimination.pdf>.

[85] Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>.

[86] BOE-A-2022-11589 Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación. Disponible en: <https://www.boe.es/buscar/pdf/2022/BOE-A-2022-11589-consolidado.pdf>.

[87] La Carta de Derechos Digitales de España, presentada el 14 de julio de 2021. Disponible en: <https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf>.

ausencia de sesgos. En Italia, la *Legge 23 settembre 2025, n.132*, exige en su artículo 11.3 garantizar la protección contra la discriminación en función de atributos protegidos, sin contemplar *proxies* o agrupaciones algorítmicas.[88] En Francia, pese a la ausencia de una ley específica, el *Défenseur des droits* (Defensor de derechos) ha señalado que este tipo de discriminación puede deberse al carácter sesgado de los datos y la utilización de criterios que pueden parecer muy alejados de los motivos prohibidos por ley, pero que los sistemas correlacionan para obtener resultados adversos equivalentes (uso de *proxies*).[89] En conjunto, la comprensión del problema en estos países es parcial y, dado que se encuentran sujetos a la RIA, heredan sus falencias conceptuales.

Esta misma limitación se observa en el plano supranacional, donde la Organización de las Naciones Unidas (ONU) ha reiterado en distintos pronunciamientos la necesidad de políticas públicas que permitan afrontar los efectos desfavorables que los sistemas automatizados generan en la equidad. Por un lado, resolución A/78/L.49 subraya la importancia de la mitigación de sesgos en todo el ciclo de vida algorítmico, pero sin afectar desproporcionadamente el desarrollo tecnológico positivo.[90] Por otro lado, la resolución A/HRC/56/68 se pronuncia sobre las formas contemporáneas de racismo y xenofobia derivadas de la implementación de inteligencia artificial, alertando el riesgo potencial de la reproducción de sesgos, la generación de variables *proxy* y la formación de bucles de retroalimentación.[91] Pese a su pertinencia, los pronunciamientos de la ONU adolecen de la misma carencia que la RIA, en tanto no integran sus preocupaciones dentro de una sola construcción conceptual; circunstancia preocupante, puesto que diversos países los adoptan como directrices base para su desarrollo legislativo.[92]

Si bien las fuentes normativas citadas no son erradas *per se*, a la luz de lo desarrollado en la sección 2, puede argumentarse que estas resultan limitadas e incompletas desde una perspectiva descriptiva integral. Por este motivo, es necesario establecer un punto de apoyo conceptual que logre capturar la complejidad real de las dinámicas adversas involucradas en el fenómeno estudiado. Bajo esta perspectiva, se formulará una *descripción material* de la discriminación algorítmica, utilizando el adjetivo *material* para hacer referencia a la manera, es decir *el cómo*, los sistemas algorítmicos pueden afectar negativamente el principio de no discriminación en la realidad observable.

La descripción material propuesta en la presente investigación es la siguiente: La discriminación algorítmica se configura cuando un algoritmo, que ha sido implementado para la realización de tareas predictivas o la toma de decisiones automatizadas, genera resultados que restringen o privan el acceso justo a bienes, servicios u oportunidades a un individuo o grupo en base a correlaciones que: incorporan sesgos en perjuicio de atributos protegidos; replican patrones

[88] *Legge 23 settembre 2025, n.132*, *«Disposizioni e deleghe al Governo in materia di intelligenza artificiale»*. Disponible en: <https://www.promopa.it/wp-content/uploads/2025/09/legge_132_2025.pdf>.

[89] DÉFENSEUR DES DROITS, «Lutter contre les discriminations produites par les algorithmes et l'IA», 2024, PP. 1-5. Cabe precisar que, en la fecha de la última actualización de esta investigación (12 de octubre de 2025), Francia no ha promulgado una ley específica sobre discriminación algorítmica. Por ello, se ha considerado pertinente aludir esta ficha técnica emitida por el *Défenseur des droits* (Defensor de derechos), donde se ofrecen directrices prácticas para enfrentar los riesgos de la inteligencia artificial y que tienen el potencial de orientar futuras políticas legislativas en el país. Descarga disponible en: <https://www.defenseurdesdroits.fr/sites/default/files/2024-02/FICHE7_AlgoIA_0.pdf>.

[90] Resolución A/78/L.49, del 11 de marzo de 2024, «Aprovechar las oportunidades de sistemas seguros, protegidos y fiables de inteligencia artificial para el desarrollo sostenible». Documento disponible para su descarga en: <https://documents.un.org/doc/undoc/ltd/n24/065/95/pdf/n2406595.pdf>.

[91] Resolución A/HRC/56/68, del 3 de junio de 2024, «Contemporary forms of racism, racial discrimination, xenophobia and related intolerance». Este documento ofrece una cobertura más completa del problema. Disponible para su descarga: <https://documents.un.org/doc/undoc/gen/g24/084/20/pdf/g2408420.pdf>.

[92] Por ejemplo, la resolución A/78/L.49 ha servido como base para la promulgación de la *resolução nº 32, de 09 de outubro de 2024* en Brasil, la cual dispone sobre el enfrentamiento de la discriminación racial en relación con la inteligencia artificial. Disponible en: <https://www.gov.br/participamaisbrasil/blob/baixar/59773>.

discriminatorios utilizando variables sustitutas; o generan agrupaciones algorítmicas que funcionan como inmutables artificialmente y presentan una irrelevancia moral que no justifica dicha restricción.

La propuesta establecida, a la que se referirá en adelante como *P*, no pretende abordar el dilema jurídico sobre si este tipo de discriminación encaja dentro de las categorías de discriminación directa o discriminación indirecta, ni su posible correspondencia en las doctrinas del *disparate treatment* o *disparate impact* desarrolladas en el derecho estadounidense. Tampoco tiene por finalidad esclarecer cuestiones en materia de responsabilidad civil o penal, sin embargo, en la sección 3.1 se sugerirán algunos lineamientos que podrían orientar futuras investigaciones en la materia. Además, P no pretende ser una propuesta legislativa, sino una herramienta doctrinal que sirva como punto de partida para el debate académico pertinente. Entendido esto, a continuación, se desarrollará una explicación detallada sobre P fraccionándola en tres partes; la implementación, la restricción y las correlaciones.

3.1. Sobre la implementación

En la primera parte de P, se ha considerado que la discriminación algorítmica se configura como producto de la implementación de un algoritmo en la realización de tareas predictivas o la toma de decisiones automatizadas. Por un lado, al reflexionar sobre el proceso de implementación se deben contemplar los aspectos esenciales respecto a quién emplea el modelo, cómo opera dicha utilización y en qué contextos es problemático aplicarlo, asunto que se desarrollará en la sección 3.2. Por otro lado, es menester sustentar el por qué se considera que la realización de tareas predictivas o la toma de decisiones automatizadas son, en específico, las formas de utilización que generan los efectos negativos estudiados.

En primer lugar, la decisión sobre el despliegue de un modelo como herramienta para la ejecución de tareas recae exclusivamente en los agentes humanos. El implementador puede ser de naturaleza privada, en el caso de personas naturales o jurídicas; o de naturaleza pública, cuando un ente gubernamental recurre a la utilización de estos sistemas para aumentar la eficiencia en las funciones públicas. En este sentido, P parte de la premisa según la cual el modelo que genera impactos discriminatorios es aquel cuyo despliegue ha sido puesto en funcionamiento por un sujeto «responsable del despliegue» que, conforme a lo establecido en el artículo 3.4 del RIA, es quien «utiliza un sistema bajo su propia autoridad, salvo cuando su uso se enmarque en una actividad personal de carácter no profesional».

En lo que respecta al modo de operatividad del algoritmo cuando es implementado, se ha adoptado la clasificación expuesta en las secciones 2.4 y 2.5, donde se distingue entre modelos con interacción estática o dinámica. En atención a ello, el análisis de lo propuesto debe contemplar los posibles escenarios adversos derivados del desajuste y la ocurrencia de bucles de retroalimentación negativa. Asimismo, conviene subrayar que P delimita la discriminación algorítmica como aquella que emerge de los resultados derivados del funcionamiento del modelo, excluyendo los supuestos donde el actor humano instrumentaliza dolosamente el algoritmo con el fin de reproducir efectos discriminatorios. Dicho escenario constituye, en rigor, una manifestación de discriminación humana tradicional, pero mediada por un sistema tecnológico complejo. No obstante, la ausencia de acción directa del agente humano en los resultados no exime de responsabilidad al implementador, quien debe velar por el mantenimiento, actualización y auditoria de esta herramienta que tiene el potencial de generar impactos lesivos al derecho fundamental de la igualdad.

En segundo lugar, en P se plantea que la realización de tareas predictivas y la toma decisiones automatizadas son, en concreto, las modalidades de utilización que generan los efectos negativos analizados. Según Araujo et al., los algoritmos implementados para procesos decisionales pueden

funcionar bajo dos esquemas: 1) como instrumento de apoyo en la toma de decisiones humana, ofreciendo recomendaciones predictivas que orientan a los usuarios en determinada dirección, y, 2) como agente decisor completamente automatizado, ejecutando decisiones en nombre de instituciones u organizaciones sin requerir la participación humana en el proceso.[93]

En lo que respecta a las recomendaciones predictivas, Binns advierte que, aunque los sistemas ayudan a evitar las posibles inconsistencias de los decisores humanos, las variables empleadas por los modelos para predecir pueden contener estructuras discriminatorias que, al ofrecerse como resultados confiables, son utilizados para tomar decisiones que repercuten en el desarrollo de vida de las personas.[94] Del mismo modo, la automatización decisional puede exponer a los destinatarios de la decisión a perjuicios asociados a los resultados discriminatorios desarrollados a lo largo del presente artículo.[95] Este riesgo se ve acentuado por la capacidad del modelo para enmascarar los resultados erróneos con una aparente neutralidad, limitando de esta manera a rendición de cuentas y dispersando el foco de responsabilidad, en tanto la decisión se percibe como desvinculada de cualquier agente humano identificable.[96] En términos sencillos, es el proceso decisional, en cualquiera de sus dos dimensiones, el que expone a los individuos a efectos derivados de la ejecución de resultados afectados por patrones discriminatorios, generando desigualdades injustificadas en el acceso a recursos esenciales.

3.2. Sobre la restricción:

En la segunda parte de P, se sostiene que la discriminación algorítmica surge cuando los resultados generados por el sistema restringen o privan el acceso justo a bienes, servicios y oportunidades a un individuo o grupo. En tal planteamiento, el termino *justicia* es entendido como sinónimo de *equidad* en su dimensión distributiva, orientada a evitar que en el proceso algorítmico decisional se originen predicciones asimétricas que lesionen el principio de igualdad y no discriminación. No obstante, conviene resaltar que el concepto de equidad presenta múltiples nociones profundamente heterogéneas, lo cual obstaculiza la satisfacción conjunta de todas ellas en el diseño y desarrollo de los sistemas automatizados.[97] En este sentido, teniendo en cuenta que no todo resultado diferenciador es discriminatorio, se ha considerado que se produce un efecto injusto cuando las decisiones generadas privan o restringen, o recomiendan privar o restringir, la igualdad de condiciones en el acceso a recursos esenciales para la autonomía y el libre desarrollo humano.

Sobre este aspecto, en P se ha establecido como recursos esenciales los bienes, servicios y oportunidades. En primer lugar, la utilización del vocablo *bien* no alude únicamente a elementos físicos o tangibles, sino que abarca de igual manera los derechos y libertades, los cuales constituyen bienes jurídicos fundamentales. En este sentido, los sistemas pueden restringir de manera discriminatoria el correcto goce de derechos como el debido proceso,[98] la presunción de

[93] ARAUJO ET AL., «In AI we trust? Perceptions about automated decision making by artificial intelligence», *AI & Society*, 2020, p. 613.
[94] BINNS, «Human Judgment in Algorithmic Loops: Individual Justice and Automated Decision-Making», *Regulation & Governance*, 2020, p. 204..
[95] MÖKANDER ET AL., «Ethics Based Auditing of Automated Decision Making Systems: Nature, Scope, and Limitations», *Science and Engineering Ethics*, 2021, p. 2.
[96] BAROCAS ET AL., *op. cit.*, p. 29.
[97] PESSACH/SHMUELI, «Algorithmic Fairness», 2020, p. 5; STARKE ET AL., «Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature», *Big Data & Society*, 2022, p. 4.
[98] GREEN, «The Flaws of Policies Requiring Human Oversight of Government Algorithms», *Computer Law & Security Review*, 2022, pp. 1-22.

inocencia,[99] derechos laborales,[100] entre otros. En segundo lugar, la incorporación de algoritmos que replican patrones negativos puede traducirse en la privación injustificada del acceso a *servicios* o en su prestación con estándares de calidad reducidos para ciertos grupos poblacionales. Una manifestación concreta de esta dinámica fue descrita en la sección 2.3, donde se tomó como referencia el sistema de crédito social chino, el cual restringe el acceso de determinados medios de trasporte a personas etiquetadas como *desacreditadas* dentro de una lista negra. En tercer lugar, la obtención de *oportunidades* puede ser condicionada de manera adversa por modelos que operan asignando rangos de puntuación a individuos, como a pacientes en estado crítico, candidatos a un puesto de trabajo, postulantes a una beca o solicitantes de un crédito hipotecario.

En este sentido, algunos contextos problemáticos que merecen especial atención en relación con la limitación de recursos esenciales son los siguientes: La identificación biométrica, que conlleva riesgos significativos para derechos vinculados a la privacidad y la no discriminación.[101] Además, puede agravar desigualdades en perjuicio de personas con discapacidad u otra condición que dificulte el uso de tecnologías biométricas comunes.[102] El control fronterizo, donde los sistemas tienden a exacerbar daños xenófobos en la gestión migratoria, generando decisiones sesgadas respecto a solicitudes de residencia o visado.[103] El ámbito educativo, en el que puede reproducirse sesgos en procesos como la admisión, la evaluación académica y el diseño curricular.[104] El área laboral, en particular cuando los modelos son utilizados para el filtrado de currículos, el reclutamiento y la gestión de recursos humanos.[105] Los servicios financieros, donde el proceso decisional automatizado puede catalizar la segregación económica en el puntaje crediticio, los préstamos hipotecarios y la segmentación de clientes.[106] El sector salud, en el cual la generación de desigualdades representa una amenaza latente para cierto grupo de pacientes, lo que compromete el acceso pleno a beneficios terapéuticos, atención médica de calidad y el consentimiento informado de decisiones clínicas.[107]

3.3. Sobre las correlaciones:

En términos sencillos, la *correlación* es una métrica estadística que permite cuantificar el grado de asociación entre dos o más variables y que los algoritmos utilizan para generar resultados infiriendo patrones existentes en los datos. Por ejemplo, un modelo puede determinar la fecha propicia para el lanzamiento de una bebida energética al mercado a partir de la proximidad que

[99] BLOUNT, «Applying The Presumption Of Innocence To Policing With AI», 2021, pp. 33-48.
[100] KELLY-LYTH, «Algorithmic Discrimination at Work», *European Labour Law Journal*, vol. 14(2), 2023, pp. 152-171.
[101] KIESLICH/LÜNICH, «Regulating AI-Based Remote Biometric Identification. Investigating the Public Demand for Bans, Audits, and Public Database Registrations», 2024, pp. 1-19.
[102] RATHGEB ET AL., «Demographic Fairness in Biometric Systems: What Do the Experts Say?», 2023, pp. 1-9.
[103] SAUNDERS, «Security, digital border technologies, and immigration admissions: Challenges of and to non-discrimination, liberty and equality», *European Journal of Political Theory*, 2023, p. 7; TOMASEV ET AL., «Manifestations of xenophobia in AI systems», *AI & Society*, vol. 40, 2025, p. 745.
[104] CHINTA ET AL., «FairAIED: Navigating Fairness, Bias, and Ethics in Educational AI Applications», 2024, pp, 1-47.
[105] ALBAROUDI ET AL., «A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring», *AI*, vol. 5, 2024, pp. 383-404; KÖCHLING/WEHNER, «Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development», *Business Research*, vol. 13, 2020, pp. 795-848.
[106] BAJRACHARYA ET AL., «Recent Advances in Algorithmic Biases and Fairness in Financial Services: A Survey», 2022, pp. 1-15.
[107] CROSS ET AL., «Bias in Medical AI: Implications for Clinical Decision-Making», *PLOS Digital Health*, vol. 3(11), 2024, pp. 1-19.

esta variable tiene con factores como los eventos deportivos, condiciones climáticas, calendarios festivos, ingreso promedio de la zona, entre otros. En la tercera parte de P, se designa que los resultados algorítmicos que dan paso a consecuencias discriminatorias son aquellos basados en correlaciones que incorporan sesgos en perjuicio de atributos protegidos; replican patrones discriminatorios utilizando variables sustitutas; o generan agrupaciones algorítmicas que funcionan como inmutables artificialmente y presentan una irrelevancia moral que no justifica la restricción de recursos esenciales. Aunque se ha profundizado anteriormente sobre cada una de las asociaciones negativas mencionadas, conviene especificar lo siguiente.

Primero, las correlaciones identificadas por los algoritmos pueden contener sesgos estructurales derivados de los datos de entrenamiento, la arquitectura subyacente del modelo o la forma de interacción que los usuarios tienen con dicha herramienta. Indistintamente del supuesto que se analice, en P se ha considerado como efecto discriminatorio la utilización de asociaciones sesgadas como base decisional en procesos automatizados, generando desventajas significativas para individuos pertenecientes a grupos protegidos. Segundo, en P se identifican como resultados discriminatorios aquellos que resultan de correlaciones basadas en el empleo de variables aparentemente neutrales que, al funcionar como sustitutos de atributos protegidos, replican los mismos efectos excluyentes de la variable original. Tercero, si bien la teoría de Wachter[108] sobre la inmutabilidad artificial y los grupos algorítmicos ha sido formulada recientemente, en P se ha determinado que el efecto perjudicial generado por tales agrupaciones es funcionalmente equiparable al provocado por las correlaciones sesgadas y las variables sustitutas, por lo que resulta necesaria su inclusión en la descripción planteada. Cuarto, los fenómenos discriminatorios desarrollados en las secciones 2.4 y 2.5 han sido excluidos del análisis sobre las correlaciones, dado que P los reconoce como riesgos asociados al proceso de implementación.

4. Conclusión

Los algoritmos poseen la capacidad de generar formas de discriminación con un nivel de complejidad mayor a la observada en humanos, en tanto operan encontrando patrones estadísticos que resultan opacos, inestables, cambiantes, ambiguos e inexplicables. Bajo esta perspectiva, se advierte que los modelos pueden reproducir sesgos en perjuicio de características tuteladas por el derecho antidiscriminatorio y utilizar variables sustitutas de estas para replicar el mismo patrón negativo en contra de tales grupos vulnerables. A ello se le suma que, debido a su alta capacidad de procesamiento, los sistemas tienen el potencial de crear agrupaciones que funcionan como criterio base para la ejecución de limitaciones que repercuten en la autonomía y el proyecto de vida de las personas. Además, los efectos adversos pueden verse exacerbados como resultado de los aspectos técnicos de la implementación, en particular cuando se manifiestan desajustes algorítmicos tras un despliegue estático o bucles de retroalimentación negativos producto del despliegue dinámico.

En un sentido elemental, la discriminación algorítmica ocurre cuando un individuo o grupo de individuos ve condicionado injustificadamente (en base a características protegidas, sustitutos de estas o agrupaciones algorítmicas) su acceso a bienes, servicios u oportunidades como resultado de las predicciones generadas por un modelo que ha sido implementado en procesos decisionales en forma de recomendaciones o de manera completamente automatizada. No obstante, el marco legal vigente ha tratado este efecto adverso con enfoques disgregados, sin

[108] WACHTER, *op. cit.*, pp. 1-50.

consolidar una conceptualización unívoca que refleje las múltiples preocupaciones que este fenómeno genera en la sociedad y que la propia normativa a reconocido en sus diversos pronunciamientos. Por ello, se puede concluir que es necesaria una *descripción material de la discriminación algorítmica* que funja como base doctrinal para, por un lado, permitir la comprensión integral del problema en base a la evidencia empírica, y, por el otro, encause el debate jurídico desde un enfoque amplio que garantice la protección de los derechos fundamentales ante los riesgos del acelerado desarrollo tecnológico.

Entendido esto, el enfoque legislativo no debe orientarse en frenar la proliferación o despliegue de los algoritmos, sino que debe enfocarse en garantizar que estos sean sometidos a mecanismos de mantenimiento, actualización y auditoría continua, con el fin de evitar que su funcionamiento perpetue patrones negativos que lesionen el derecho a la igualdad y no discriminación. De igual forma, como se ha evidenciado a lo largo de la presente investigación, el marco doctrinal actual debe considerar el carácter sociotécnico de la problemática analizada, por lo que algunos conceptos tradicionales podrían requerir una reformulación o ampliación que permita la integración de aspectos técnicos no contemplados en el análisis jurídico convencional. Además, dada la dificultad que representa la construcción de modelos que satisfagan simultáneamente todas las posibles definiciones de justicia y equidad, conviene establecer un consenso descriptivo del problema, a fin de que se tutelen los derechos fundamentales de manera efectiva.

5. Bibliografía

ADAM, George et al., «Hidden Risks of Machine Learning Applied to Healthcare: Unintended Feedback Loops Between Models and Future Data Causing Model Degradation», *Proceedings of Machine Learning Research*, 2020, pp. 1-22.

ADAMS-PRASSL, Jeremias et al., «Directly Discriminatory Algorithms», *The Modern Law Review*, 2023, pp. 144-175.

ALBAROUDI, Elham et al., «A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring», *AI*, vol. 5, 2024, pp. 383-404.

ALELYANI, Salem, «Detection and Evaluation of Machine Learning Bias», *Applied Sciences*, 2021, pp. 1-17.

ALFAIF, Yousef, «Recommender Systems Applications: Data Sources, Features, and Challenges», *Information*, 2024, pp. 1-25.

ANDRINGA, Sible/GODFROID, Aline, «Sampling Bias and the Problem of Generalizability in Applied Linguistics», *Annual Review of Applied Linguistics*, 2020, pp. 134-142.

ARAUJO, Theo et al., «In AI we trust? Perceptions about automated decision making by artificial intelligence», *AI & Society*, 2020, pp. 611-623.

BAJRACHARYA, Aakriti et al., «Recent Advances in Algorithmic Biases and Fairness in Financial Services: A Survey», 2022, pp. 1-15.

BAROCAS, Solon et al., «Fairness and Machine Learning. Limitations and Opportunities», 2023, pp. 1-284.

BAUER, Kevin et al., «Feedback Loops in Machine Learning: A Study on the Interplay of Continuous Updating and Human Discrimination», *Journal of the Association for Information Systems*, vol. 25(4), 2024, pp. 804-866.

BINNS, Reuben, «Human Judgment in Algorithmic Loops: Individual Justice and Automated Decision-Making», *Regulation & Governance*, 2020, pp. 197-211.

BLOUNT, Kelly, «Applying The Presumption Of Innocence To Policing With AI», 2021, pp. 33-48.

BOLSON RUZZARIN, Mateus, «Conversation with Noam Chomsky - The Responsibility of Intellectuals», 2021. Disponible en: https://www.youtube.com/watch?v=7M35aasejgI&t=78s

BORGSTEDE, Matthias/EGGERT, Frank, «Squaring the Circle: From Latent Variables to Theory-Based Measurement», *Theory & Psychology*, 2023, pp. 118-137.

BRIGHT, Jonathan et al.. «Generative AI is Already Widespread in the Public Sector», 2024, pp. 1-10.

CABIDDU, Franccesca et al., «Why do Users Trust Algorithms? A Review and Conceptualization of Initial Trust and Trust Over Time», *European Management Journal*, vol. 40, núm. 5, 2022, pp. 685-706.

CHAKRABORTY, Joymallya et al., «Bias in Machine Learning Software: Why? How? What to Do?», Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '21), 2021, pp. 429-440.

CHEN, Mo/GROSSKLAGS, Jens, «Social Control in the Digital Transformation of Society: A Case Study of the Chinese Social Credit System», *Social Sciences*, vol. 11, 2022, pp. 1-23.

CHEN, Zhenpeng, «A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers», *ACM Transactions on Software Engineering and Methodology*, 2023, pp. 1-31.

CHINTA, Sribala et al., «FairAIED: Navigating Fairness, Bias, and Ethics in Educational AI Applications», 2024, pp, 1-47.

CIRCIUMARU, Alexandru, «Futureproofing EU Law. The Case of Algorithmic Discrimination», *University of Oxford,* 2021, pp. 1-123.

CROSS, James, «Bias in Medical AI: Implications for Clinical Decision-Making», *PLOS Digital Health*, vol. 3(11), 2024, pp. 1-19.

DASTIN, Jeffrey, «Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women», *Reuters*, 2018. Obtenido de: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

DAVIES, Benjamin/DOUGLAS, Thomas, «Learning to Discriminate: The Perfect Proxy Problem in Artificially Intelligent Sentencing», *Sentencing and Artificial Intelligence*, 2022, pp. 97-121.

DÉFENSEUR DES DROITS, «Lutter contre les discriminations produites par les algorithmes et l'IA», 2024, PP. 1-5. Obtenido de: https://www.defenseurdesdroits.fr/sites/default/files/2024-02/FICHE7_AlgoIA_0.pdf

DOLATA, Mateusz et al., «A Sociotechnical View of Algorithmic Fairness», *Information Systems Journal*, 2021, pp. 1-87.

ENGSTROM, David et al., «Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies», 2020, pp. 1-122.
FARIČ, Ana/BRATKO, Ivan , «Machine Bias: A Survey of Issues», *Informatica*, núm. 48, 2024, pp. 205–212.

FORAN, Michael, «Grounding Unlawful Discrimination», *Legal Theory*, vol. 28, 2022, pp. 3-34.

FORD, Christopher et al., «Racial Differences in Performance of HbA1c for the Classification of Diabetes and Prediabetes among US Adults of Non-Hispanic Black and White Race», *U.S. Department of Health and Human Services*, vol. 36(10), 2019, pp. 1234-1242.

GAMA, Ricardo et al., «Multi-Agent Environments for Vehicle Routing Problems», 2024., pp. 1-17.

GARBELLINI FILHO, Luiz, «El enfrentamiento a la discriminación interseccional en el sistema interamericano de derechos humano. Análisis de las aportaciones del marco de la OEA a la construcción del derecho discriminatorio», en FILLOL MAZO, Adriana (coord.)*, Los logros de la gobernabilidad en América Latina*, Dykinson S.L., 2024, pp. 127-148.

GHAFFARY, Shirin, «How TikTok's Hate Speech Detection Tool Set Off a Debate About Racial Bias on the App», *Vox,* 2021. https://www.vox.com/recode/2021/7/7/22566017/tiktok-black-creators-ziggi-tyler-debate-about-black-lives-matter-racial-bias-social-media

GRANT, Nico/HILL, Kashmir, «Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's», *The New York Times,* 2023. https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html

GREEN, Ben, «The Flaws of Policies Requiring Human Oversight of Government Algorithms», *Computer Law & Security Review*, 2022, pp. 1-22.

HAN, Xudong et al., «Balancing out Bias: Achieving Fairness Through Balanced Training», 2022, pp. 1-16.

HASAN, Emrul et al., «Review-based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives», *Proceedings of the ACM Measurement Anaysis of Computing Systems*, 2024, pp. 1-33.

HEAVEN, Will, «Predictive Policing Algorithms Are Racist. They Need to Be Dismantled», *MIT Technology Review*, 2020. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

HINDER, Fabian et al., «One or Two Things We Know About Concept Drift. A Survey on Monitoring Evolving Environments», 2023, pp. 1-44.

HIREVUE, «About Us», *Hirevue*, 2025. Obtenido de: https://www.hirevue.com/about

HOVAKIMYAN, Gurgen/BRAVO, Jorge, «Evolving Strategies in Machine Learning: A Systematic Review of Concept Drift Detection», *Information*, 2024, pp. 1-24.

HU, Lily, «What is "Race" in Algorithmic Discrimination on the Basis of Race?», *Journal of Moral Philosophy*, 2023, pp. 1-23.

IMANA, Basileal et al., «Auditing for Discrimination in Algorithms Delivering Job Ads», *Proceedings of the Web Conference 2021 (WWW '21)*, 2021, pp. 3767-3778.

INNERARITY, Daniel, «Justicia Algorítmica y Autodeterminación Deliberativa», *Isegoría Revista de Filosofía moral y política*, núm. 68, 2023, p. 1-10.

IRIONDO, Roberto, «Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women», *Carnegie Mellon University*, 2018. Obtenido de: https://ml.cmu.edu/news/news-archive/2018/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women

KELLY-LYTH, Aislinn, «Algorithmic Discrimination at Work», *European Labour Law Journal*, vol. 14(2), 2023, pp. 152-171.

KIESLICH, Kimon/LÜNICH, Marco, «Regulating AI-Based Remote Biometric Identification. Investigating the Public Demand for Bans, Audits, and Public Database Registrations», 2024, pp. 1-19.

KING, OWEN/MERTENS, Mayli, «Self Fulfilling Prophecy in Practical and Automated Prediction», *Ethical Theory and Moral Practice*, 2023, pp. 134-135.

KÖCHLING, Alina/WEHNER, Marius, «Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development», *Business Research*, vol. 13, 2020, pp. 795-848.

LANDA ARROYO, Cesar, «El Derecho Fundamental a la Igualdad y No Discriminación en la Jurisprudencia del Tribunal Constitucional del Perú», *Estudios Constitucionales*, vol. 19, núm. 12, 2021, pp. 86-87.

MALIK, Esraa et al., «Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture», *Mathematics*, 2022, pp, 1-16.

MATA SIERRA, María, «La Discriminación por Indiferenciación y su Incidencia en el Ámbito Tributario», *Revista Jurídica de la Universidad de León*, núm. 8, 2021, pp. 185-201.

MEHRABI, Ninareh et al., «A Survey on Bias and Fairness in Machine Learning», *ACM Computing Surveys (CSUR)*, 2022, pp. 1-35.

MÖKANDER, Jakob et al, «Ethics Based Auditing of Automated Decision Making Systems: Nature, Scope, and Limitations», *Science and Engineering Ethics*, 2021, pp. 1-30.

MUÑOZ, GUTIÉRREZ, Catherine, «La Discriminación en una Sociedad Automatizada: Contribuciones desde América Latina», *Revista Chilena de Derecho y Tecnología*, 2021, p. 271-307.

NITTLE, Nadra, «Spend "Frivolously" and Be Penalized Under China's New Social Credit System», *Vox,* 2018. Obtenido de: https://www.vox.com/the-goods/2018/11/2/18057450/china-social-credit-score-spend-frivolously-video-games

NOWAK, Marcin, «The Impact of Rule Based Decision Engines on Business Efficiency», *Higson*, 2024. Obtenido de: https://www.higson.io/blog/the-impact-of-rule-based-decision-engines-on-business-efficiency

PÁEZ, Andrés, «Negligent Algorithmic Discrimination», *Law and Contemporary Problems*, 2021, pp. 19-33.

PAGAN, Nicolò et al., «A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems», 2023, pp. 1-2

PAPADOPOULOS, Petros et al. «A Systematic Review of Technologies and Standards Used in the Development of Rule Based Clinical Decision Support Systems», *Health and Technology*, 2022, pp. 713-727.

PAVLIDIS, Georgios, «Unlocking the Black Box: Analysing the EU Artificial Intelligence Act's Framework for Explainability in AI», *Law Innovation and Technology*, vol. 16, núm. 1, 2024, pp. 293-308.

PESSACH, Dana/ SHMUELI, Erez, «Algorithmic Fairness», 2020, pp. 1-31.

PLATKIN, Matthew/ IYER, Sundeep, «Guidance on Algorithmic Discrimination and the New Jersey Law Against Discrimination», 2025, p. 10.

PRINCE, Anya/SCHWARCZ, Daniel, «Proxy Discrimination in the Age of Artificial Intelligence and Big Data», *Iowa Law Review*, 2020, pp. 1257-1318.

RAMÍREZ-BUSTAMANTE, Natalia/PÁEZ, Andrés,. «Análisis Jurídico de la Discriminación Algorítmica en los Procesos de Selección Laboral», 2021, pp. 1-29.

RATHGEB, Christian et al., «Demographic Fairness in Biometric Systems: What Do the Experts Say?», 2023, pp. 1-9.

RAVISHANKAR, Pavan et al., «Provable Detection of Propagating Sampling Bias in Prediction Models», *The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)*, 2023, pp. 9562-9569.

ROA AVELLA, Marcela del Pilar et al., «Uso del Algoritmo COMPAS en el Proceso Penal y los Riesgos a los Derechos Humanos», *RevIsta Brasileira de Direito Processual Penal*, 2022, pp. 275-310.

SALOMÉ RESURRECCIÓN, liliana, «La discriminación y algunos de sus calificativos: directa, indirecta, por indiferenciación, interseccional (o múltiple) y estructural», *Pensamiento Constitucional*, 2017, pp. 261.

SAUNDERS, Natasha, «Security, digital border technologies, and immigration admissions: Challenges of and to non-discrimination, liberty and equality», *European Journal of Political Theory*, 2023, p. 7

SCHWARTING, Rena/ULBRICHT, Lena. «Why Organization Matters in "Algorithmic Discrimination"», *Köln Z Soziol*, 2022, pp. 307-330.

SHAHBAZI, Nima, «Representation Bias in Data: A Survey on Identification and Resolution Techniques», *Woodstock '18: ACM Symposium on Neural Gaze Detection,* 2021, pp. 1-47.

SINGLA, Alex et al., «The state of AI in early 2024: Gen AI adoption spikes and starts to generate value», *QuantumBlack AI by McKinsey*, 2024, pp. 1-22

SÓLMUNDSDÓTTIR, Agnes et al., «Mean Machine Translations: On Gender Bias in Icelandic Machine Translations», *Proceedings of the 13th Conference on Language Resources and Evaluation*, 2022, pp. 3113-3121.

STARKE, Christopher et al., «Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature», *Big Data & Society*, 2022, pp. 1-16

STRAUB, Vincent et al., «Artificial intelligence in government: Concepts, standards, and a unified framework», *Computer Science,* 2023, pp. 1-34.

SURESH, Harini et al., «Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making», *12th ACM Conference on Web Science*, 2020, pp. 314-324.

TAPIA LARA, Francisco, «A Conceptual Framework for Simulating Feedback Loops in Engineering Design», *The University of Leeds*, 2023, pp. 1-199.

TEO, Sue Anne, «Artificial intelligence and its 'slow violence' to human rights», *AI and Ethics*, 2024, pp. 1-16.

TOMASEV, Nenad et al., «Manifestations of xenophobia in AI systems», *AI & Society*, vol. 40, 2025, pp. 741-763.

WACHTER, Sandra, «Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law», *Tulane Law Review*, 2022, pp. 1-50.

WAHID, Sumra, «How to Get Away With Discrimination: The Use of Algorithms to Discriminate in the Internet Entertainment Industry», *American University Journal of Gender, Social Policy & the Law*, 2023, pp. 107-139.

WEERTS, Hilde et al., «Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms», *2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 2024, pp. 1-11.

WEIZENBAUM-INSTITUT., «Sandra Wachter - The Theory of Artificial Immutability», 2023. Disponible en: https://www.youtube.com/watch?v=khw1aXupTgk

WÖRLE, Franziska, «Negative Feedback Loops and Self-fulfilling Prophecies: Sociotechnical Assessment of Unfairness in Predictive Algorithms», *Department of Computer Science Ludwig-Maximilians-Universität at München*, 2024, pp. 1-129.

WYLLIE, Sierra et al., «Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias», 2024, pp. 1-46.

YUAN, Haochen et al., «Your Offline Policy is Not Trustworthy: Bilevel Reinforcement Learning for Sequential Portfolio Optimization», 2025, pp. 1-21.