# Sesgos algorítmicos en los sistemas de inteligencia artificial implementados en administración pública. Un análisis de casos sobre modelos de predicción de riesgo

# Vieses algorítmicos em sistemas de inteligência artificial implementados na administração pública. Uma análise de caso sobre modelos de previsão de risco

# *Algorithmic Biases in Artificial Intelligence Systems Implemented in Public Administration. A Case Analysis of Risk Prediction Models*

**Eugenio Arguelles Toache** *

El uso de la inteligencia artificial (IA) en la administración pública ha crecido significativamente en la última década debido a los múltiples beneficios de esta tecnología en las funciones de gobierno. Sin embargo, esto también ha dejado en evidencia una serie de efectos adversos como la presencia de sesgos algorítmicos que derivan en decisiones que pueden ser discriminatorias e injustas para ciertos individuos y grupos sociales. El objetivo de esta investigación es identificar y analizar las causas de los sesgos algorítmicos presentes en los sistemas de IA de la administración pública, así como sus consecuencias éticas, políticas y sociales. Para ello se analizan cuatro casos representativos a nivel internacional de modelos de predicción de riesgo basados en IA que son utilizados en diferentes ámbitos de la administración pública. El resultado de este análisis muestra que los sesgos algorítmicos se producen en la etapa de diseño y de entrenamiento de los algoritmos, generando afectaciones importantes en el acceso a derechos fundamentales, servicios públicos y programas sociales. La sociedad civil juega un papel fundamental al momento de evidenciar los sesgos algorítmicos y sus consecuencias, así como para contribuir a generar estrategias para la implementación ética de esta tecnología.

**Palabras clave**: modelos de predicción de riesgos; sesgos algorítmicos; discriminación algorítmica; justicia algorítmica; gobierno electrónico

* Becario posdoctoral SECIHTI adscrito al Instituto de Investigaciones Sociales de la Universidad Nacional Autónoma de México (IIS-UNAM). Doctor en ciencias sociales en el área de economía y gestión de la innovación por la Universidad Autónoma Metropolitana (UAM) de México. Correo electrónico: eugenio.toache@gmail.com. ORCID: https://orcid.org/0000-0002-0121-1681.

A utilização da inteligência artificial (IA) na administração pública tem crescido significativamente na última década devido aos múltiplos benefícios desta tecnologia nas funções governamentais. No entanto, isto também revelou uma série de efeitos adversos, como a presença de enviesamentos algorítmicos que geram resultados e decisões que podem ser discriminatórias e injustas para alguns indivíduos ou grupos sociais. O objetivo desta investigação é identificar e analisar as causas dos enviesamentos algorítmicos presentes nos sistemas de IA da administração pública, bem como as suas consequências éticas, políticas e sociais. Para tal, são analisados quatro casos internacionalmente representativos de modelos de previsão de risco baseados em IA, utilizados em diferentes áreas da administração pública. Os resultados desta análise mostram que os enviesamentos algorítmicos ocorrem nas etapas de conceção e treino dos algoritmos, gerando impactos significativos no acesso a direitos fundamentais, serviços públicos e programas sociais. A sociedade civil desempenha um papel fundamental na exposição dos enviesamentos algorítmicos e das suas consequências, bem como na contribuição para a geração de estratégias para a implementação ética desta tecnologia.

**Palavras-chave:** modelos de previsão de risco; enviesamentos algorítmicos; discriminação algorítmica; justiça algorítmica; governo eletrônico

*The use of artificial intelligence (AI) in public administration has grown significantly in the last decade due to the multiple benefits of this technology in government functions. However, this has also revealed a series of adverse effects such as the presence of algorithmic biases that generate results and decisions that can be discriminatory and unfair for some individuals or social groups. The objective of this article is to identify and analyze the causes of algorithmic biases present in AI systems in public administration, as well as their ethical, political, and social consequences. To do so, four internationally representative cases of AI-based risk prediction models that are used in different areas of public administration are analyzed. The result of this analysis shows that algorithmic biases occur at the design and training stage of the algorithms, generating significant impacts on access to fundamental rights, public services, and social programs. Civil society plays a fundamental role in exposing algorithmic biases and their consequences, as well as in contributing to the generation of strategies for the ethical implementation of this technology.*

***Keywords**: risk prediction models; algorithmic biases; algorithmic discrimination; algorithmic justice; e-government*

## Introducción

En la última década, se ha evidenciado un crecimiento significativo en el uso de la inteligencia artificial (IA) por parte de la administración pública. Este crecimiento se aceleró considerablemente tras la pandemia de COVID-19, marcando un punto de inflexión en la adopción de tecnologías digitales para la gestión pública. La IA está transformando la forma en que los gobiernos operan, enfocándose en la optimización de procesos administrativos internos, la formulación más precisa de políticas públicas, la toma de decisiones basada en datos, la mejora en la prestación de servicios públicos y la creación de canales de comunicación más eficientes con la ciudadanía. Las principales áreas en las que la administración pública ha implementado IA incluyen la salud pública, la seguridad y el orden público, el transporte público y la gestión de la movilidad, la prevención y la gestión de desastres naturales, la asistencia social y la administración tributaria.

No obstante, el uso de la IA en el ámbito gubernamental ha suscitado un extenso debate en la academia y la sociedad en general sobre sus posibles beneficios y los potenciales efectos adversos. Entre los efectos adversos más discutidos destacan los sesgos algorítmicos, un fenómeno ampliamente documentado en la literatura especializada. Por ejemplo, Valle-Cruz *et al.* (2024) llevaron a cabo una revisión sistemática de investigaciones previas y concluyeron que los sesgos algorítmicos representan uno de los impactos negativos más frecuentes y preocupantes del uso de la IA en la administración pública. Los sesgos algorítmicos son tendencias o desviaciones que influyen en las valoraciones, estimaciones o predicciones generadas por los algoritmos y que afectan de manera desproporcionada a determinados individuos o grupos sociales. La pregunta general que guía esta investigación es: ¿cuáles son las causas y consecuencias de los sesgos algorítmicos en los sistemas de IA utilizados por la administración pública y cuáles son las estrategias más utilizadas para eliminar o mitigar estos sesgos y sus consecuencias?



Los sesgos algorítmicos provienen principalmente de dos fuentes. En primer lugar, los sesgos algorítmicos se pueden producir en el proceso de diseño cuando se toman decisiones que reflejan consciente o inconscientemente los perjuicios de quienes los desarrollan o cuando se toman decisiones erróneas que generan desviaciones en los parámetros y reglas que conforman el algoritmo. En segundo lugar, los sesgos algorítmicos se pueden producir en el proceso de entrenamiento cuando los algoritmos son entrenados con datos históricos que reflejan prejuicios sociales o desigualdades sistémicas, datos que no son representativos o que han sido erróneamente etiquetados. Como resultado, los sistemas algorítmicos no solo replican estas desigualdades y estos perjuicios, sino que, debido a su capacidad de automatizar y escalar procesos, pueden amplificarlas a gran escala.

Los sesgos presentes en los algoritmos que guían decisiones gubernamentales tienen el potencial de provocar efectos adversos significativos, como la perpetuación o amplificación de dinámicas de exclusión y discriminación estructural. En concreto,

los sesgos algorítmicos pueden traducirse en políticas públicas y decisiones administrativas que marginen sistémicamente a ciertos grupos sociales en función de características como su nacionalidad, origen étnico, género, idioma o religión. Estas dinámicas discriminatorias no solo refuerzan desigualdades históricas, sino que también obstaculizan el acceso equitativo a derechos fundamentales, servicios públicos esenciales y programas sociales diseñados para mejorar la calidad de vida de la población. Ante esta realidad, el análisis de los sesgos algorítmicos, más que un tema técnico, es un desafío ético y político que exige la responsabilidad de los gobiernos para evitar que la IA refuerce desigualdades. Para ello es esencial establecer marcos regulatorios sólidos, realizar auditorías independientes, fomentar la transparencia y garantizar la participación de diversos actores sociales. Abordar estos sesgos no solo es una cuestión de justicia social, sino también de legitimidad institucional, asegurando que la IA contribuya a la equidad y el bienestar colectivo en lugar de perpetuar la discriminación.

El objetivo de este artículo de investigación es examinar una selección de casos representativos a nivel internacional para identificar y analizar los sesgos algorítmicos presentes en los modelos de predicción de riesgos basados en IA utilizados en la administración pública, enfocando el análisis en tres aspectos principales: i) las causas de estos sesgos vinculadas tanto al diseño de los algoritmos como a los datos empleados en su entrenamiento; ii) sus consecuencias en términos de sus afectaciones sobre el acceso a derechos fundamentales, servicios públicos y programas sociales; y iii) las principales estrategias utilizadas para abordar y mitigar dichos sesgos, destacando las mejores prácticas y lecciones aprendidas de cada caso estudiado.



Los cuatro casos que se analizarán son: COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), Algoritmo AMS (*Arbeits Markt Service*) y SyRI (*System Risk Indication*) y la Plataforma Tecnológica de Intervención Social. La documentación de cada caso se realizará a través del análisis de diversas fuentes, incluyendo artículos científicos, reportajes periodísticos, documentos institucionales y contenidos disponibles en sitios web. Previo a eso, el siguiente apartado ofrece una revisión y discusión de la literatura científica sobre los sesgos algorítmicos.

## 1. Sesgos algorítmicos: definición, causas y consecuencias

De acuerdo con Simon *et al*. (2020), los sesgos algorítmicos se refieren a errores recurrentes en sistemas computacionales que, de manera injusta, discriminan a ciertos individuos o grupos sobre otros al asignarles resultados desfavorables basados en criterios inapropiados o irracionales y perpetuando las desigualdades presentes en la sociedad. Pasquinelli *et al.* (2022) complementan esta idea y observan que los sesgos algorítmicos son, en esencia, una amplificación de prejuicios preexistentes en los datos y en la incorporación inadvertida de prejuicios humanos en la estructura y las reglas de los algoritmos, lo cual ocurre debido a errores computacionales, una comprensión inadecuada de la información o deficiencias en las técnicas de aproximación utilizadas

en el aprendizaje automático. En síntesis, los sesgos algorítmicos son manifestaciones de patrones, tendencias y errores que subyacen en los algoritmos y que conducen a estimaciones, predicciones y valoraciones desiguales que favorecen o perjudican a ciertos individuos o grupos sociales, generando resultados que, en última instancia, son discriminatorios, excluyentes e injustos.

Los sesgos algorítmicos, según CAF (2021) y Martín (2022), tienen su origen esencialmente en dos fuentes principales: el diseño de los algoritmos y los datos utilizados para entrenarlos. Los sesgos en el proceso de diseño de los algoritmos se deben a que estos son diseñados por seres humanos que toman decisiones sobre las reglas y los parámetros que definen la estructura de los algoritmos, de modo que los sesgos algorítmicos surgen de estas decisiones. En primer lugar, los seres humanos pueden, de manera involuntaria, incorporar sus propias experiencias, valores, perspectivas del mundo y prejuicios durante el proceso de diseño de algoritmos, por lo que dichas influencias se manifestarán en la configuración de parámetros, la definición de reglas y la selección de criterios de evaluación, lo que da lugar a modelos con tendencias y desviaciones inherentes (Martín, 2022). En segundo lugar, los diseñadores de los algoritmos pueden tomar, de manera involuntaria, decisiones erróneas sobre las variables, métricas y parámetros que componen a los algoritmos, generando igualmente tendencias y desviaciones en los modelos (CAF, 2021). En tercer lugar, aunque la mayoría de estos sesgos se introduce de manera involuntaria, también pueden ser añadidos intencionalmente para manipular o engañar a ciertos grupos de individuos, lo que plantea implicaciones éticas graves (Sued, 2022). De esta manera, los algoritmos dejan de ser neutrales y reproducen las subjetividades de sus creadores, influyendo en cómo se interpretan y procesan las necesidades y características de diferentes grupos sociales



Por otro lado, los sesgos algorítmicos también emergen de las características de los datos utilizados para entrenar los modelos algorítmicos y sus procesos de aprendizaje automático. En primer lugar, estos datos pueden contener sesgos históricos, es decir, prejuicios acumulados a lo largo del tiempo como resultado de decisiones humanas y prácticas pasadas, estos sesgos se perpetúan y amplifican cuando los algoritmos los replican sin cuestionarlos (Baker & Hawn, 2022; Koniakou, 2023). En segundo lugar, los datos utilizados para entrenar a los algoritmos pueden ser incompletos o no representativos de ciertas poblaciones, lo que lleva a que el algoritmo produzca generalizaciones erróneas y resultados sesgados (Kordzadeh & Ghasemaghaei, 2022). Finalmente, incluso si los datos son completos y representativos y no contienen sesgos históricos, las decisiones humanas tomadas durante el proceso de clasificación y etiquetado, como la selección, eliminación y priorización de algunas variables y datos, pueden introducir sesgos inadvertidos (CAF, 2021)

En resumen, los sesgos algorítmicos son una extensión de los prejuicios humanos y pueden surgir en distintas etapas del ciclo de vida de un algoritmo. Aparecen en el diseño, cuando los desarrolladores incorporan reglas sesgadas, en el entrenamiento, al usar datos incompletos o con prejuicios históricos, y en la evaluación, al validar los

resultados con métricas sesgadas (CAF, 2021). Aunque en términos generales los sesgos algorítmicos son el reflejo de los sesgos humanos, sus consecuencias pueden ser significativamente más graves y de mayor alcance. Esto se debe a que la IA se encuentra en un proceso de adopción masiva y crecimiento exponencial tanto en la industria como en la administración pública; por lo tanto, existe el riesgo de amplificar y perpetuar estos sesgos a una escala sin precedentes. Es decir, el uso intensivo y generalizado de la IA permite que los sesgos algorítmicos no solo se proyecten sobre una gran cantidad de personas, sino que también extiendan sus efectos adversos en múltiples contextos y a lo largo del tiempo (Martín, 2022).

Los sesgos algorítmicos generan una representación desproporcionada de ciertos grupos sociales frente a otros, lo que puede ocurrir de manera sistémica al basarse en características como raza, género, nacionalidad o clase social. Esto implica que los algoritmos pierden la neutralidad y objetividad esperadas, ya que introducen desviaciones y tendencias que resultan en un trato desigual, así como en predicciones o valoraciones que no son equitativas entre diferentes grupos identitarios (Kordzadeh & Ghasemaghaei, 2022). Una vez que los sistemas de IA comienzan a operar, los sesgos integrados en ellos se replican constantemente, produciendo resultados excluyentes, discriminatorios e inequitativos. Estos resultados, a su vez, pueden ser utilizados como datos de entrada para entrenar nuevamente los algoritmos, creando un círculo vicioso de retroalimentación sesgada. Este proceso amplifica y perpetúa las desigualdades sociales, ya que cada interacción refuerza los prejuicios integrados en el sistema (Martín, 2022; Koniakou, 2023).



A diferencia de los sesgos humanos, que están limitados por las interacciones individuales, los sesgos algorítmicos tienen un alcance exponencial. La capacidad de los algoritmos para procesar y tomar decisiones de manera autónoma, en combinación con su uso extendido en sectores como la salud, la educación, las finanzas y la justicia, multiplica el impacto de estos sesgos, afectando a millones de personas en todo el mundo. Además, al no ser detectados o corregidos adecuadamente, estos sesgos se convierten en estructuras rígidas que perpetúan desigualdades históricas bajo una apariencia de objetividad tecnológica.

El riesgo asociado a los sesgos algorítmicos en los sistemas de IA empleados en la administración pública es especialmente preocupante, debido a la magnitud y el impacto potencial de las decisiones que estos sistemas generan. A diferencia de otros contextos, en el ámbito de la administración pública las decisiones basadas en algoritmos tienen el potencial de afectar directamente a un volumen significativo de la población durante periodos prolongados de tiempo. Esto implica que los sesgos presentes no solamente se reproducen y amplifican en una escala mayor, sino que también tienden a perpetuarse, reforzando desigualdades existentes y creando nuevas barreras estructurales (Martín, 2022).

Adicionalmente, el uso de sistemas algorítmicos en la administración pública tiene una particular trascendencia debido a las áreas estratégicas y prioritarias en las que suelen

aplicarse. Estos sistemas influyen en aspectos fundamentales para el bienestar social, como el acceso a derechos humanos, la distribución de recursos en programas sociales y la prestación de servicios públicos esenciales, como salud, educación y justicia (CAF, 2021; Koniakou, 2023). De esta manera, cuando los sesgos algorítmicos se filtran en la toma de decisiones gubernamentales, se convierten en un problema no solo técnico, sino también profundamente ético, político y de justicia social. Las decisiones injustas basadas en algoritmos pueden consolidar dinámicas de exclusión, discriminación y marginación, especialmente hacia grupos históricamente desfavorecidos. Estas dinámicas, además, suelen ser menos visibles y más difíciles de cuestionar, debido al aparente carácter objetivo y neutral de las tecnologías algorítmicas.

En esta investigación se analizan los modelos de predicción de riesgos basados en IA, los cuales analizan grandes volúmenes de datos para identificar patrones y estimar la probabilidad de que ocurran ciertos eventos adversos. En la administración pública, estos modelos se utilizan para optimizar la toma de decisiones en áreas como seguridad, salud, bienestar social y justicia. Por ejemplo, pueden evaluar el riesgo de reincidencia delictiva, predecir la vulnerabilidad de ciertos grupos o detectar posibles fraudes en programas sociales.

La hipótesis de esta investigación es que los sesgos algorítmicos en los modelos de predicción de riesgo utilizados en la administración pública se originan tanto en el diseño del algoritmo, debido a la selección sesgada de variables predictivas, como en su entrenamiento, al emplear datos que reflejan desigualdades estructurales. Estos sesgos pueden llevar a la sobreestimación del riesgo en grupos vulnerables, reforzando patrones de discriminación y exclusión, lo que impacta el acceso equitativo a derechos, programas sociales y servicios públicos. En el siguiente apartado se describe la metodología utilizada para realizar el análisis de las causas y consecuencias de los sesgos algorítmicos presentes en los sistemas de IA utilizados en la administración pública.



## 2. Metodología

Para la selección de los casos de estudio, se empleó la base de datos del Observatorio de Algoritmos con Impacto Social (OASI, por sus siglas en inglés) de Éticas Foundation, la cual recopila y analiza 152 casos a nivel global en los que el uso de algoritmos de IA ha generado consecuencias sociales negativas no previstas. El propósito de esta base de datos es visibilizar la problemática y fomentar la conciencia sobre la necesidad de auditar y regular el uso de algoritmos en ámbitos que impactan directamente a la sociedad. De los 152 casos documentados, 96 corresponden a algoritmos implementados en la administración pública y, tras un primer análisis exploratorio, se identificaron 24 casos específicamente relacionados con modelos algorítmicos diseñados para la predicción de riesgos. Posteriormente, estos 24 casos fueron examinados puntualmente, priorizando aquellos que contaban con suficiente información detallada para un análisis riguroso y priorizando la diversidad de los casos

en cuanto a su ámbito de implementación y los países de origen. Como resultado de este proceso de selección, se eligieron cuatro casos para su estudio exhaustivo, permitiendo así una evaluación más detallada de las causas de los sesgos algorítmicos y de sus impactos políticos y sociales.

Los casos seleccionados son los siguientes: i) Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), algoritmo utilizado en el sistema judicial estatal de los Estados Unidos para predecir el riesgo de reincidencia delictiva; ii) System Risk Indication (SyRI), algoritmo implementado en Países Bajos para predecir el riesgo de fraude en solicitudes de subsidios y apoyos sociales; iii) Allegheny Family Screening Tool (AFST), algoritmo implementado en el condado de Allegheny en Pensilvania, Estados Unidos, para predecir el riesgo de negligencia y abuso infantil; y iv) Plataforma Tecnológica de Intervención Social, algoritmo implementado en la provincia de Salta, Argentina, para predecir el riesgo de deserción escolar y de embarazo en adolescentes.

Para recopilar información sobre los casos de estudio, se empleó un enfoque basado en el análisis documental y la revisión de sitios web. La documentación se sustentó en tres tipos de fuentes principales: artículos científicos, documentos y reportes institucionales, y fuentes hemerográficas. La obtención de estos documentos se llevó a cabo mediante la búsqueda del nombre exacto de cada caso en bases de datos académicas como Scopus y Google Scholar, así como en el motor de búsqueda de Google. Con el objetivo de garantizar la rigurosidad del análisis, se estableció un criterio de prioridad en el uso de las fuentes, dando preferencia a los artículos científicos, seguidos de los reportes institucionales y, en última instancia, de las fuentes hemerográficas. En cuanto al análisis de sitios web, se realizó una búsqueda de las páginas oficiales de los proyectos estudiados o de las organizaciones responsables de su desarrollo e implementación, con el propósito de obtener información directa y verificada sobre cada caso. Adicionalmente, para fortalecer el análisis de los casos estudiados, se incluyen ejemplos de casos similares implementados en otros países, lo que permiten contextualizar y enriquecer los hallazgos obtenidos.



## 3. Análisis de casos

### 3.1. Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)

El sistema COMPAS es uno de los modelos de predicción de riesgo basados en IA más conocidos y fue desarrollado por la empresa Northpointe en el año 2000. Este sistema es ampliamente utilizado en el sistema judicial de los Estados Unidos para predecir el riesgo de reincidencia delictiva y apoyar el proceso de toma de decisiones de los jueces en torno a la concesión de libertad bajo fianza antes del juicio o de la libertad condicional de los condenados, una vez que han cumplido una parte de la sentencia. Mediante un algoritmo de IA, COMPAS analiza hasta 12 variables personales, como

la edad, los antecedentes penales y los problemas de adicciones o educación, para asignar a cada detenido o condenado una puntuación de riesgo del 1 al 10. Estas puntuaciones se agrupan en tres niveles: bajo nivel de reincidencia (1-4), medio nivel de reincidencia (5-7) y alto nivel de reincidencia (8-10). Aquellos clasificados como de alto riesgo tienen muy pocas posibilidades de obtener libertad bajo fianza o libertad condicional; aquellos clasificados como de riesgo medio pueden acceder a este beneficio con condiciones de vigilancia estricta; mientras que aquellos clasificados como de bajo riesgo suelen recibir este beneficio con menores restricciones (Blomberg *et al*., 2010).

El sistema COMPAS se ha convertido en un caso emblemático y ampliamente citado de sesgos algorítmicos en la administración pública. En 2016, gracias a una investigación periodística de ProPublica, se descubrió la existencia de un sesgo algorítmico racial en este sistema. La investigación de Angwin *et al*. (2016) analizó 11.757 casos que fueron evaluados mediante COMPAS y encontró que este sistema clasificaba de manera desproporcionada a personas de piel negra con un riesgo más alto de reincidencia en comparación con personas de piel blanca; en concreto, las personas de piel negra fueron clasificadas erróneamente como de alto riesgo casi el doble de veces que las personas de piel blanca que no reincidieron (45% frente a 23%). Incluso ajustando por factores como delitos previos, edad y género, los acusados de piel negra tenían un 45% más de probabilidades de recibir puntuaciones más altas de riesgo general y un 77% más de probabilidades de alto riesgo de reincidencia violenta.



En menos de 18 meses, más de 200 artículos académicos citaron el artículo de ProPublica y emplearon métodos alternativos para reproducir el análisis llegando a resultados similares y a la misma conclusión: el sistema COMPAS presenta un sesgo algorítmico racial (Washington, 2018). Por ejemplo, la investigación de Hamilton (2019) utilizó la misma base de datos que la investigación de ProPublica y encontró que COMPAS también presenta un sesgo algorítmico hacia las personas hispanas al asignarles clasificaciones que sobrestiman su riesgo de reincidencia.

Autores como Fuchs (2018), Avella *et al*. (2022) y Marinucci *et al.* (2023) señalan que los sesgos algorítmicos presentes en el sistema COMPAS tienen su origen en el uso de datos históricos marcados por prejuicios, desigualdades estructurales y discriminación sistémica en el sistema legal, lo cual se trasfiere al algoritmo mediante el proceso de entrenamiento. Estos datos incluyen patrones desproporcionados de informes de delitos provenientes de vecindarios predominantemente habitados por personas afroamericanas e hispanas, reflejando las desigualdades y los prejuicios sociales que afectan a estas comunidades. Según Avella *et al.* (2022), las comunidades históricamente sometidas a una vigilancia y persecución desproporcionadas por parte de las fuerzas del orden, en particular las de bajos ingresos y las minorías raciales, enfrentan ahora un nuevo nivel de discriminación, ya que el algoritmo tiende a calificarlas con puntuaciones más altas de reincidencia. Este fenómeno perpetúa y amplifica las desigualdades existentes, transformando las disparidades humanas en decisiones algorítmicas con consecuencias significativas para la equidad y la justicia.

Adicionalmente, estudios recientes han identificado que el sistema COMPAS también exhibe un sesgo algorítmico relacionado con el género, el cual tiende a sobreestimar el riesgo de reincidencia en mujeres. Este sesgo se debe a un error en el proceso de diseño del algoritmo, ya que fue diseñado para ser neutral en términos de género y evitar prejuicios; sin embargo, termina generando el efecto contrario. Al no considerar el género como una variable predictiva, evalúa el riesgo de reincidencia de manera uniforme para hombres y mujeres, pese a que las mujeres presentan tasas significativamente menores de reincidencia. Este sesgo incrementa la probabilidad de que sean tratadas de manera injusta en el ámbito de la justicia penal, al enfrentarse a decisiones como la negación de la libertad bajo fianza o condicional, o sean objeto de niveles de supervisión excesivos en caso de obtener dichos beneficios, afectando con ello sus derechos fundamentales (Skeem *et al*., 2016; Hamilton, 2018). Para mitigar este sesgo de género en el sistema COMPAS, sus desarrolladores han implementado recientemente normas específicas para hombres y mujeres, ajustando la clasificación de las puntuaciones según el género. No obstante, la adopción de estas normas depende de los funcionarios judiciales de cada Estado o condado, quienes en su mayoría han rechazado su aplicación, argumentando que la neutralidad de género debe prevalecer al tratarse de un grupo protegido que no debería influir en las decisiones de justicia penal (Hamilton, 2018).

Como consecuencia de los sesgos algorítmicos presentes en el sistema COMPAS, no solamente se reproducen las desigualdades estructurales presentes en la sociedad, sino que, al ser utilizado para tomar decisiones judiciales clave, como otorgar libertad provisional o condicional, las perpetúa y amplifica. Esto afecta la equidad en el acceso a beneficios judiciales para las personas de piel negra, otras minorías y mujeres, resultando especialmente afectadas las mujeres afroamericanas, erosionando la confianza en la imparcialidad del sistema de justicia. Este ciclo de retroalimentación entre datos sesgados y decisiones basadas en ellos agrava las barreras estructurales existentes y contribuye a mantener dinámicas de discriminación institucionalizada (CAF, 2021).



A pesar de las evidencias que señalan la existencia de sesgos algorítmicos en el sistema COMPAS, la empresa Northpointe ha rechazado estas afirmaciones, argumentando que su herramienta es precisa para predecir la reincidencia general, los actos violentos y la incomparecencia ante el tribunal. Además, el algoritmo de COMPAS está protegido por leyes de secreto comercial, lo que impide que se conozca su funcionamiento exacto y dificulta el análisis y la deliberación en contextos legales. Esta falta de transparencia, sumada a la falta de un examen exhaustivo de las implicaciones constitucionales por parte de la Corte Suprema de los Estados Unidos, ha dejado el tema sin una resolución definitiva, ya que las discusiones en tribunales estatales y federales inferiores no han logrado abordar plenamente la problemática.

Alrededor del mundo existen casos similares al sistema COMPAS que igualmente han sido objeto de críticas por la presencia de sesgos algorítmicos. Por ejemplo, en Cataluña, España, se utiliza la herramienta RisCanvi (acrónimo de las palabras

"riesgo" y "cambio" en catalán) y -aunque la única investigación independiente que existe demuestra que esta herramienta no presenta sesgos algorítmicos de género, raza, nacionalidad o edad- Aróstegui (2023) indica que el uso de datos de entrenamiento desactualizados y no representativos puede generar predicciones de riesgo sesgadas sobre los grupos menos representados en las prisiones de Cataluña, como las mujeres y los extranjeros.

### 3.2. System Risk Indication (SyRI)

En 2012, el gobierno de Países Bajos diseñó e implementó el modelo de predicción de riesgo conocido como SyRI con el propósito de elaborar indicadores de riesgo para detectar posibles casos de fraude en solicitudes de subsidios y apoyos sociales. Este sistema se basaba en un algoritmo predictivo de IA que procesaba una amplia cantidad de datos personales de los ciudadanos, abarcando aspectos como empleo, sanciones administrativas, información fiscal, propiedades, beneficios sociales, antecedentes educativos, integración cívica, cumplimiento normativo, deudas, pensiones y permisos (Berning, 2023). Posteriormente, el algoritmo comparaba estos datos con perfiles de riesgo construidos a partir de información de personas previamente identificadas como defraudadoras; al identificar similitudes generaba una lista de individuos o entidades con alta probabilidad de incurrir en fraude relacionado con las prestaciones sociales ofrecidas por el gobierno (R3D, 2020). Finalmente, este análisis daba lugar a la elaboración de informes de riesgo que servían para investigar a posibles defraudadores de manera más detallada, lo cual podía derivar en su exclusión o suspensión de los programas sociales, o en acusaciones de fraude (Lazcoz & Castillo, 2020).



En octubre de 2018, un relator especial de la ONU sobre pobreza extrema y derechos humanos remitió un informe al Tribunal de La Haya en el que señalaba serias deficiencias en el funcionamiento del algoritmo SyRI, principalmente la presencia de sesgos algorítmicos que generaban un efecto discriminatorio y estigmatizador hacia los ciudadanos pertenecientes a determinados grupos étnicos y con un menor nivel de ingresos (R3D, 2020; Lazcoz & Castillo, 2020). En posteriores investigaciones se descubrió que los sesgos étnicos se originaron debido a que el algoritmo fue diseñado para clasificar las nacionalidades diferentes a la holandesa como un indicador de riesgo, reflejando una discriminación implícita en su programación (Peeters & Widlak, 2023). Este sesgo fue exacerbado por el uso de datos históricos para entrenar el algoritmo, datos que ya contenían prejuicios estructurales y actitudes discriminatorias contra minorías étnicas, principalmente aquellas con raíces en las antiguas colonias holandesas de Surinam o en las islas caribeñas holandesas, así como grupos con herencia en Turquía y Marruecos (Arts & Van den Berg, 2024). Igualmente, los sesgos socioeconómicos se presentaron debido a que los datos utilizados para entrenar al algoritmo reflejaban discriminación hacia personas que habitaban zonas caracterizadas por sus bajos ingresos (Berning, 2023). Como resultado, SyRI perpetuó y amplificó estas desigualdades estructurales al generar clasificaciones que asociaban desproporcionadamente un mayor riesgo a las personas de nacionalidades

diferentes, los grupos étnicos minoritarios y las personas con bajos ingresos. La falta de auditorías externas e independientes para evaluar los efectos adversos de este sistema permitió que los resultados discriminatorios se perpetuaran y retroalimentaran a lo largo del tiempo.

Los sesgos algorítmicos presentes en el sistema SyRI y sus resultados discriminatorios tuvieron un severo impacto en la vida de miles de personas. De acuerdo con Peeters y Widlak (2023), entre 2012 y 2019, más de 30.000 ciudadanos fueron injustamente acusados de cometer fraude, lo que desencadenó en la pérdida de beneficios sociales fundamentales, la exclusión de programas de asistencia gubernamental y una grave situación de endeudamiento, ya que muchos de ellos fueron obligados a reembolsar todos los beneficios económicos que habían recibido hasta el momento. Para las personas afectadas, en su mayoría provenientes de comunidades migrantes o de bajos ingresos, las consecuencias de estas acciones resultaron devastadoras y, en muchos casos, irreparables. Estas medidas no solo las empujaron a situaciones de pobreza extrema y, en algunos casos, a la bancarrota, sino que también afectaron profundamente múltiples aspectos de sus vidas; por ejemplo, muchas personas informaron haber desarrollado problemas de salud mental, como depresión y estrés crónico, y otras sufrieron el estigma social de ser consideradas defraudadoras afectando permanentemente sus relaciones personales y sociales, mientras que las madres y padres de más de 2000 infantes perdieron su custodia como consecuencia de estas decisiones discriminatorias, lo cual no solo desintegró familias, sino que también generó un trauma intergeneracional que amplificó la vulnerabilidad de los menores y limitó severamente sus oportunidades de un futuro estable (Arts & Van den Berg, 2024).



En 2020, el Tribunal de Distrito de La Haya declaró ilegal el uso del algoritmo SyRI, sentando un precedente significativo en Europa, ya que se trató de la primera sentencia para declarar ilegal un algoritmo en dicho continente. El fallo señaló que este sistema violaba derechos fundamentales, como la privacidad y la protección contra la discriminación, aunado a la falta de transparencia del algoritmo y su incapacidad para evitar sesgos. El caso SyRI desencadenó una crisis política de gran magnitud que provocó la renuncia del tercer gabinete del gobierno del primer ministro Mark Rutte en 2021, mientras que en 2022 el gobierno holandés reconoció formalmente que este sistema y las decisiones derivadas fueron un ejemplo de racismo institucional. A pesar de esto, las consecuencias para las familias afectadas persisten, y muchas de ellas continúan enfrentando largas batallas legales en busca de justicia y de las indemnizaciones correspondientes (Arts & Van den Berg, 2024).

En 2016 se presentó un caso similar a SyRI en Australia cuando se implementó el sistema RoboDebt, un algoritmo de IA diseñado para detectar posibles pagos excesivos en asistencia social mediante la comparación de los ingresos reportados con los subsidios recibidos, si detectaba alguna discrepancia se emitía automáticamente avisos de cobro de deudas a los ciudadanos. Sin embargo, el sistema presentaba un sesgo socioeconómico que afectó principalmente a los estratos socioeconómicos más

bajos, generando errores en el cálculo de sus ingresos y provocando interrupciones en los pagos de asistencia y graves problemas de endeudamiento para miles de ciudadanos, ya que se emitieron indebidamente 470.000 deudas (Rinta-Kahila *et al*., 2022). Tras una investigación y críticas generalizadas, el gobierno australiano eliminó el sistema, enfrentó una demanda colectiva y reembolsó alrededor de 1200 millones de dólares a los afectados. Igualmente, el Departamento de Trabajo y Pensiones (DWP, por sus siglas en inglés) del Reino Unido, implementó en 2016 un algoritmo de predicción de riesgos para detectar posibles fraudes en los beneficios sociales. Sin embargo, en 2021 se reportó que dicho algoritmo presentaba un sesgo hacia las personas con discapacidad, clasificándolas con un mayor riesgo de cometer fraude y resultando en investigaciones intrusivas y humillantes para dichas personas. Esto se tradujo en acciones legales por parte de los afectados para que se diera a conocer el funcionamiento del algoritmo (Savage, 2021).

### 3.3. Allegheny Family Screening Tool (AFST)

En 2016, el Departamento de Servicios Humanos del condado de Allegheny, en Pensilvania, Estados Unidos, implementó el modelo predictivo de riesgo conocido como AFST para predecir e identificar posibles casos de negligencia y abuso infantil con el objetivo de que los trabajadores sociales tomen decisiones más informadas y realicen las intervenciones necesarias para garantizar la seguridad y el bienestar de los infantes. Esta herramienta utiliza un algoritmo de IA para procesar y analizar datos provenientes de diferentes fuentes (Chouldechova *et al*., 2018; Blanchard, 2022): i) llamadas telefónicas relacionadas con posibles casos de negligencia y abuso infantil realizadas por miembros de la comunidad al Departamento de Servicios Humanos; ii) datos históricos de informes que contienen denuncias de maltrato previas o denuncias por abuso de drogas o alcohol; y iii) registros administrativos relevantes como ser beneficiario de algún programa social, recibir tratamiento de salud mental, estar en un programa de libertad condicional o habitar en un vecindario de bajos ingresos. A partir de estas variables predictivas, el algoritmo genera una puntuación de riesgo para cada caso, en una escala del 1 al 20, donde 20 representa el nivel más alto de riesgo. Esta puntuación tiene como objetivo priorizar las acciones de los trabajadores sociales en función del riesgo calculado; por ejemplo, los casos con puntaciones de riesgo bajo generalmente son descartados o archivados, y los casos con puntuaciones de riesgo medio no requieren una intervención inmediata, por lo que los trabajadores sociales establecen un monitoreo continuo o inician una investigación más exhaustiva, mientras que los casos con puntuaciones de riesgo alto requieren una intervención inmediata por parte de los trabajadores sociales: realizar visitas al hogar, aplicación de entrevistas, asesoramiento psicológico, incorporación a programas de intervención familiar o –en el caso más grave– separación de niños de sus familias de origen y posterior asignación a una organización de acogida (Gerchick *et al*., 2023; Rittenhouse *et al*., 2023).



Poco después de su implementación, la herramienta AFST comenzó a recibir críticas de diversos investigadores debido a la presencia de sesgos algorítmicos, particularmente raciales y socioeconómicos. Por ejemplo, las investigaciones de

Eubanks (2017) y Chouldechova *et al.* (2018) revelaron que la herramienta asigna puntuaciones de riesgo desproporcionadamente altas a familias de color y en situación de pobreza. Los sesgos socioeconómicos surgen debido a que el algoritmo está diseñado para utilizar variables predictivas vinculadas directamente con la pobreza, como el uso de beneficios gubernamentales, programas sociales, servicios de salud pública, no tener suficiente comida, tener una vivienda inadecuada o dejar a un niño solo mientras se trabaja (Eubanks, 2017). Aunque estas variables buscan identificar posibles factores asociados con la negligencia o el abuso infantil, terminan reflejando y perpetuando desigualdades estructurales. Adicionalmente, el uso de datos históricamente sesgados en el entrenamiento del algoritmo agrava esta problemática, ya que el Departamento de Servicios Humanos interviene con mayor frecuencia en comunidades de bajos recursos, por lo que se generan más informes sobre estas poblaciones, los cuales asocian históricamente la pobreza con la negligencia y el abuso infantil (Rittenhouse *et al*., 2023). Todo esto se traduce en que el algoritmo establezca una correlación errónea entre la pobreza y la negligencia infantil, lo que resulta en una mayor probabilidad de que estas familias sean clasificadas como de alto riesgo, independientemente de las condiciones reales del entorno familiar.

Por otro lado, los sesgos raciales se originan por la alta cantidad de familias de color y multirraciales en las bases de datos utilizadas por el Departamento de Servicios Humanos. En primer lugar, los denunciantes de la comunidad llaman a las líneas de abuso y negligencia infantil para reportar frecuentemente casos que involucran a familias negras y multirraciales tres veces y media más veces en comparación con los casos relacionados con familias blancas, generando con ello más informes y sobrerrepresentación (Eubanks, 2017). En segundo lugar, los registros administrativos utilizados para alimentar el algoritmo incluyen más información sobre estos grupos raciales en comparación con otros, ya que es más probable que esos grupos participen en programas gubernamentales, lo que refuerza esta sobrerrepresentación (Chouldechova *et al*., 2018). Este patrón de denuncias y de representatividad en los datos refleja prejuicios raciales sistémicos presentes en la sociedad, los cuales se ven amplificados y perpetuados al integrarse en los datos utilizados por el algoritmo.



Estos sesgos algorítmicos generan que las familias de bajos ingresos y de color tengan una mayor probabilidad de ser clasificadas erróneamente como de alto riesgo, lo que resulta en un elevado número de falsos positivos que desencadenan investigaciones y medidas de intervención desproporcionadas, muchas veces discriminatorias e injustas, que pueden llevar incluso a la pérdida de la custodia de los infantes. Además, las puntuaciones obtenidas a través del AFST, junto con las acciones subsecuentes, pueden perpetuar el impacto negativo en estas familias, agravando los efectos de la discriminación sistémica, ya que no solamente limita las oportunidades de defensa y apelación frente a dichas decisiones, sino que también restringe el acceso a otros beneficios sociales, afectando profundamente sus derechos y su capacidad de superar las condiciones que inicialmente las colocaron en una situación vulnerable (Gerchick *et al*., 2023).

Debido a la evidencia sobre las predicciones erróneas del modelo ASFT y sus consecuencias, se decidió diseñar una segunda versión en la cual se excluyeron algunas variables relacionadas con los beneficios públicos que recibían las familias, esto con el objetivo de eliminar sesgos socioeconómicos y raciales que estaban incluidos en dichos registros administrativos (Chouldechova *et al*., 2018). Sin embargo, aunque estas modificaciones representaron algunas mejoras, en la actualidad la herramienta sigue siendo objeto de preocupaciones y críticas debido a que no ha logrado eliminar completamente los sesgos, ya que se siguen utilizando datos como el historial de tratamiento de salud mental, registros educativos y las denuncias relacionadas con el consumo de drogas o alcohol, datos que representan en forma desproporcionada a los grupos marginados, por lo que las conclusiones algorítmicas continúan reflejando las antiguas calificaciones sociales de las familias por parte de la comunidad, las autoridades y los tribunales (Blanchard, 2022).

En septiembre de 2020, el mismo condado de Allegheny implementó un modelo de predicción de riesgos similar al AFST, llamado Hello Baby, con el objetivo de evaluar el riesgo de negligencia y abuso infantil de los recién nacidos. Sin embargo, la investigación de Blanchard (2022) demostró que esta herramienta igualmente presenta sesgos algorítmicos hacia las familias de bajos ingresos y de color, ya que se basa en datos que son tan discriminatorios y sesgados como los datos utilizados por el algoritmo AFST, lo cual genera que muchas familias sean erróneamente investigadas e intervenidas inclusive antes de que ocurra el maltrato o antes de que los recién nacidos abandonen el hospital.



En otros países también se utilizan modelos de predicción de riesgos basados en IA para predecir la negligencia y el abuso infantil; por ejemplo, en 2012 se implementó la herramienta Vulnerable Children Predictive Risk Model en Nueva Zelanda, en 2014 se implementó Early Help Profiling System en el Condado en Londres y en 2018 se implementó el modelo Gladsaxe en Dinamarca. Tal y como describe Lappalainen (2021), estas herramientas también presentan sesgos algorítmicos importantes. En Nueva Zelanda la herramienta fue cuestionada y suspendida porque discriminaba a las familias maoríes y generaba una tasa desproporcionada de separaciones de niños; en Londres se detuvo la implementación del modelo debido a preocupaciones sobre su precisión y falta de transparencia; y en Dinamarca también se detuvo su implementación por cuestiones relacionadas con la vigilancia excesiva en los barrios marginales.

### 3.4. Plataforma Tecnológica de Intervención Social

En 2017 el gobierno de la provincia de Salta, Argentina, anunció un convenio con empresa Microsoft y la Fundación CONIN para la implementación de la Plataforma Tecnológica de Intervención Social, que consiste en un modelo de predicción de riesgo basado en un algoritmo de IA que pronostica el riesgo de deserción escolar y de embarazo en adolescentes. A partir de múltiples variables –asistencias, desempeño académico, origen étnico, situación de discapacidad, número de personas en el hogar, ingresos familiares, nivel educativo del entorno familiar, acceso a servicios básicos

como agua potable y saneamiento, historial de violencia de género intrafamiliar, acceso a servicios de salud sexual y reproductiva–, el algoritmo calcula perfiles de riesgo y clasifica a las adolescentes en diferentes niveles de probabilidad de deserción escolar y embarazo temprano (Pedace *et al*., 2023; Hagerty *et al*., 2023). De acuerdo con sus creadores, los resultados arrojados son utilizados para priorizar intervenciones como tutorías especializadas, apoyo emocional, estímulos económicos, servicios sociales, distribución de anticonceptivos, talleres educativos sobre salud sexual y reproductiva, y apoyo psicológico para las jóvenes identificadas (Ortiz Freuler & Iglesias, 2018).

En 2018 se realizó una prueba piloto del algoritmo en la cual se identificaron 418 niñas y niños con más de 70% de riesgo de deserción escolar y 250 niñas con más de 70% de riesgo de embarazo (Ortiz Freuler & Iglesias, 2018). Aunque en este estudio piloto se reveló que las tasas de falsos positivos fueron relativamente bajas, múltiples especialistas, académicos y organizaciones de la sociedad civil manifestaron serias preocupaciones sobre la posible reproducción de sesgos algorítmicos que pudieran ocasionar la existencia de un trato injusto y discriminatorio hacia las personas de bajos ingresos y minorías étnicas. Específicamente, el estudio de la World Web Foundation destacó la falta de transparencia del algoritmo debido a la falta de accesibilidad de las bases de datos empleadas y el proceso de diseño del modelo (Ortiz Freuler & Iglesias, 2018). Por otro lado, el Laboratorio de Inteligencia Artificial Aplicada (LIAA) de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires (UBA) analizó en profundidad el algoritmo en 2018 y publicó un informe en el que reportaron múltiples problemas como la utilización de datos sesgados e inadecuados y la reutilización del conjunto de entrenamiento como datos de evaluación, lo cual se traduce en resultados que sobredimensionaban del riesgo de deserción escolar y de embarazo en adolescentes (Hagerty *et al*., 2023).



En cuanto a los datos utilizados para entrenar y alimentar el algoritmo, Hagerty *et al*. (2023) y Pedace *et al*. (2023) señalan que estos fueron obtenidos a partir de encuestas realizadas por el gobierno provincial y diversas organizaciones de la sociedad civil. Dichas encuestas se llevaron a cabo entre 2016 y 2017 y abarcaron aproximadamente a 300.000 personas residentes en barrios de bajos ingresos, caracterizados por una alta presencia de comunidades indígenas, como los wichi, kolla y guaraní, así como de migrantes provenientes de Bolivia y Paraguay. Si bien el uso de datos socioeconómicos y demográficos es fundamental para la elaboración de modelos predictivos, su aplicación en este contexto generó preocupaciones significativas, ya que la forma en que se procesan e interpretan estos datos puede conducir a un sesgo algorítmico que resulte en una sobreestimación del riesgo de deserción escolar y embarazo adolescente entre los jóvenes de estas comunidades. Este sesgo no solamente refuerza estereotipos negativos, sino que también contribuye a la estigmatización de sectores vulnerables, al establecer una asociación implícita entre la pobreza, el abandono escolar y la maternidad temprana. Como consecuencia, se corre el riesgo de que las políticas públicas y las intervenciones basadas en estos modelos perpetúen y profundicen la desigualdad y la discriminación que sufren estas comunidades en lugar de abordar sus causas estructurales de manera efectiva.

A pesar de que los funcionarios del Ministerio de Primera Infancia defendieron la neutralidad de la base de datos, argumentando que el algoritmo no presentaba sesgos debido a que su propósito era exclusivamente la predicción dentro de poblaciones vulnerables, su implementación generó un fuerte rechazo por parte de diversas organizaciones de la sociedad civil. En particular, la predicción del embarazo adolescente fue motivo de especial controversia, ya que los colectivos feministas expresaron su oposición, señalando que el algoritmo vulneraba los derechos de niñas y mujeres al ignorar las profundas desigualdades estructurales que enfrentan. Además, alertaron sobre el riesgo de que esta herramienta pudiera ser instrumentalizada por grupos contrarios a los derechos sexuales y reproductivos para deslegitimar la necesidad de leyes que garanticen el acceso al aborto. Esta preocupación se intensificó debido a la participación de la Fundación CONIN en el desarrollo del algoritmo, una organización fundada por el pediatra Abel Albino, conocido por su activismo contra el aborto (Hagerty *et al*., 2023). Debido a estas críticas y al intenso debate público que se generó, el uso del algoritmo para predecir el riesgo de embarazo adolescente se suspendió indefinidamente, mientras que el uso del algoritmo para predecir el riesgo de deserción escolar fue suspendido hasta 2024, cuando se comenzó a implementar formalmente.

A nivel mundial, existen diversas iniciativas tecnológicas similares a la Plataforma Tecnológica de Intervención Social. Un ejemplo notable es el sistema de Alerta Temprana de Deserción Escolar (DEWS, por sus siglas en inglés), implementado en 2012 por el Estado de Wisconsin, Estados Unidos, el cual utiliza un algoritmo de IA y aprendizaje automático para predecir el riesgo de deserción escolar en estudiantes de sexto a noveno grado, permitiendo a las instituciones educativas intervenir oportunamente. Sin embargo, en 2021, un análisis realizado por el Departamento de Instrucción Pública de Wisconsin (DPI, por sus siglas en inglés) evidenció un grave sesgo algorítmico en el DEWS, el cual ocasionaba que el sistema sobreestimara significativamente el riesgo de deserción en estudiantes afroamericanos e hispanos, generando un 42% de falsos positivos en estudiantes de piel negra y un 18% en estudiantes hispanos en comparación con sus pares blancos (Feathers, 2023). Esta distorsión se debe a que el algoritmo utiliza la raza como un factor de predicción, lo que lleva a la estigmatización y discriminación de estos grupos estudiantiles históricamente desfavorecidos.



## 4. Discusión

El análisis de los casos revela que los sesgos algorítmicos presentes en los modelos de predicción de riesgos basados en IA utilizados en la administración pública tienen su origen en el proceso de diseño del algoritmo y en el proceso de entrenamiento. En la fase de diseño, los sesgos pueden surgir de manera accidental debido al tratamiento incorrecto de las variables predictivas de riesgo, o de manera deliberada cuando se seleccionan ciertas variables como predictoras de riesgo debido a los perjuicios y estereotipos presentes en los diseñadores del algoritmo; esto reproduce

las desigualdades y los procesos de exclusión preexistentes en la sociedad. El sesgo por razones de género del sistema COMPAS es un ejemplo de la introducción de un sesgo de forma accidental, ya que el algoritmo fue diseñado para ser neutral en términos de género, lo que llevó a la exclusión de esta variable predictiva. Sin embargo, al no considerar las diferencias en las tasas de reincidencia entre hombres y mujeres, el sistema terminó sobrestimando el riesgo de reincidencia en mujeres. El sesgo por razones étnicas del sistema SyRI es un ejemplo de la introducción deliberada de un sesgo, ya que los diseñadores del algoritmo consideraron que el hecho de tener una nacionalidad diferente a la holandesa era un indicador de riesgo de cometer fraude en las prestaciones sociales, lo que provocó una sobrestimación del riesgo para estos grupos.

Por otro lado, en la etapa de entrenamiento del algoritmo, los sesgos pueden surgir debido al uso de datos históricamente sesgados y a problemas de representatividad de estos datos, lo que refuerza patrones discriminatorios preexistentes, afectando la imparcialidad y la precisión de las decisiones y los resultados generados por el algoritmo. El sesgo socioeconómico del sistema AFST es un ejemplo del uso de datos de entrenamiento históricamente sesgados: debido a que el Departamento de Servicios Humanos interviene con mayor frecuencia en familias de bajos ingresos, el algoritmo se entrena con informes que históricamente han asociado la pobreza con la negligencia y el abuso infantil, lo que resulta en una sobreestimación del riesgo de estas familias. El sesgo socioeconómico de la Plataforma Tecnológica de Intervención Social ejemplifica la sobrerrepresentación de ciertos grupos en los datos de entrenamiento: debido a que las familias de bajos ingresos y las poblaciones indígenas están sobrerrepresentadas en las encuestas utilizadas para construir la base de datos, el algoritmo tiende a sobrestimar el riesgo de deserción escolar y embarazo adolescente en estos grupos.



Los sesgos introducidos en el diseño del algoritmo se refuerzan y amplifican con los sesgos introducidos en su entrenamiento; esta retroalimentación perpetúa y agrava los sesgos algorítmicos, distorsionando aún más los resultados obtenidos. Por ejemplo, el sesgo por razones étnicas del sistema SyRI surgió en el proceso de diseño al utilizar las nacionalidades distintas a la holandesa como un predictor de riesgo; sin embargo, dicho sesgo se agravó al entrenar el algoritmo con datos históricos que ya contenían prejuicios estructurales contra minorías, especialmente descendientes de antiguas colonias holandesas y comunidades de origen turco y marroquí. De esta manera, los algoritmos de los modelos para la predicción de riesgos utilizados en la administración pública pueden ser tan sesgados como las personas que los diseñaron y como los datos que se utilizan para entrenarlos.

Los casos analizados en esta investigación ponen en evidencia de manera contundente los impactos negativos de los sesgos algorítmicos en los modelos de predicción de riesgos empleados en la administración pública. Lejos de ser meros errores técnicos, estos sesgos tienen consecuencias profundas y sistémicas que afectan la vida de muchas personas, ya que no solo llevan a la sobrestimación del

riesgo de comunidades históricamente vulnerables, sino que también profundizan las desigualdades preexistentes, consolidando dinámicas de exclusión, discriminación y marginación que afectan el acceso a derechos fundamentales, programas sociales y servicios públicos esenciales. En el caso del sistema COMPAS, los sesgos algorítmicos afectan profundamente el derecho a la libertad condicional o bajo fianza de afroamericanos, hispanos y mujeres. En el caso del SyRI, los sesgos algorítmicos afectaron el acceso a programas sociales de miles de personas provenientes de comunidades migrantes y de bajos ingresos, y las condujo a situaciones severas de endeudamiento, pobreza, estigmatización social y problemas de salud mental. En el caso del AFST, los sesgos algorítmicos han ocasionado investigaciones injustas y medidas de intervención desproporcionadas hacia familias afroamericanas, multirraciales y de bajos ingresos; inclusive algunas familias han llegado a perder la custodia de sus hijas e hijos. En el caso de la Plataforma Tecnológica de Intervención Social, los sesgos algorítmicos contribuyeron a la estigmatización de miles de adolescentes provenientes de familias indígenas, migrantes y de bajos recursos.

Debido a los preocupantes efectos negativos de los sesgos algorítmicos utilizados en la administración pública, surge una cuestión fundamental que debe ser discutida a profundidad: ¿es realmente posible eliminar por completo los sesgos algorítmicos? Esta pregunta no solamente plantea un desafío técnico, sino que también abre un debate ético, político y social sobre la naturaleza misma de la IA y sus impactos en los procesos de toma de decisiones. Una respuesta factible a este cuestionamiento es que los sesgos algorítmicos en la etapa de diseño se pueden eliminar al diversificar los equipos de desarrollo, incorporando expertos de distintos grupos demográficos para reducir la influencia de perspectivas, prejuicios y sesgos de un grupo particular, promoviendo un diseño más inclusivo y equitativo en los algoritmos (Nadeem *et al*., 2022). Sin embargo, incluso con equipos de diseño con un alto grado de diversidad demográfica, los sesgos pueden persistir debido a estructuras organizacionales, normas culturales o metodologías de desarrollo que privilegian ciertos enfoques y perspectivas (Johansen *et al*., 2020). Otra respuesta factible al anterior cuestionamiento es que los sesgos algorítmicos en la etapa de entrenamiento del algoritmo pueden eliminarse al excluir información sensible de los datos de entrenamiento, como raza, género o nivel socioeconómico; sin embargo, esta estrategia no garantiza la eliminación del sesgo, ya que los algoritmos de aprendizaje automático más sofisticados pueden inferir estas características a través de variables indirectas y reconstruir implícitamente la información oculta, repicando así el sesgo original presente en los datos (Fuchs, 2018).



En este sentido, podemos afirmar que los sesgos algorítmicos son inevitables, ya que los sistemas de IA siempre estarán influenciados por valores, suposiciones, perjuicios y estereotipos de las personas u organizaciones que los crean e implementan, así como de los datos utilizados. En la práctica no existe verdaderamente un algoritmo neutral, ya que cada fase del desarrollo implica decisiones subjetivas que no solo determinan cómo el algoritmo procesa la información, sino también qué realidades prioriza y cuáles ignora. En última instancia, el algoritmo no es una verdad objetiva, sino una construcción estadística con un margen de error inherente, donde los

modelos resultantes reflejan perspectivas humanas codificadas en matemáticas y que generan resultados que son sesgados en menor o en mayor medida (Eubanks, 2017).

Dado que los sesgos son inherentes a los algoritmos, autores como Eubanks (2017), Pedace *et al*. (2023) y Aróstegui (2023) destacan la necesidad de capacitar a los profesionales que utilizan estas herramientas, argumentando que su formación no solo debe enfocarse en la identificación de sesgos, sino también en el desarrollo de habilidades para intervenir activamente en los resultados y decisiones arrojadas por los modelos. De este modo, los profesionales no solo podrán detectar y comprender las limitaciones y desviaciones de los algoritmos, sino que también estarán en condiciones de aplicar estrategias para promover un uso más transparente, ético y equitativo de ellos y mitigar sus efectos negativos. Este enfoque, conocido como *human-in-the-loop*, busca equilibrar la automatización con el juicio humano, permitiendo que los expertos evalúen, ajusten y corrijan predicciones sesgadas antes de que influyan en procesos críticos, como la asignación de recursos o la toma de decisiones en políticas públicas. En este sentido, para mitigar los efectos de los sesgos algorítmicos, es fundamental implementar mecanismos de validación humana en las decisiones automatizadas. Sin embargo, tal y como demuestran los casos analizados en esta investigación, en los cuales los algoritmos funcionan como herramientas de apoyo y sus resultados son evaluados y validados por funcionarios públicos antes de tomar decisiones, este enfoque por sí solo no siempre corrige los sesgos algorítmicos.



En primer lugar, se pueden presentar los sesgos cognitivos que se derivan de la falta de experiencia y conocimiento técnico de los funcionarios públicos y que dificultan su capacidad para evaluar críticamente las decisiones algorítmicas. Debido a esta falta de comprensión, se vuelve más difícil identificar sesgos y corregirlos, lo que perpetúa la aplicación de criterios injustos y discriminatorios (Allhutter *et al*., 2020). Esta situación se presentó en el caso del SyRI, ya que los funcionarios públicos no contaban con los conocimientos y las herramientas necesarias para realizar una auditoría algorítmica, limitando su capacidad para evaluar críticamente los resultados arrojados y ocasionando que la identificación de los sesgos ocurriera después de que el algoritmo ya había impactado negativamente a ciertos grupos (Arts & Van den Berg, 2024).

En segundo lugar, pueden darse los sesgos de automatización que indican una tendencia de los individuos a confiar de manera excesiva en los resultados generados por los algoritmos, ocasionando que los funcionarios acepten sin cuestionamientos las recomendaciones y decisiones algorítmicas, incluso cuando estas presentan sesgos evidentes. De esta manera, en lugar de aplicar un juicio propio y considerar información que contradiga la decisión del algoritmo, pueden validar automáticamente sus sugerencias, reforzando así las desigualdades subyacentes; este fenómeno también se conoce como “tecnochovinismo”: la suposición de que las computadoras son superiores a las personas, o que una solución tecnológica es superior a cualquier otra (Blanchard, 2022). Este escenario se evidenció en el caso del AFST, ya que a pesar de que los funcionarios públicos están capacitados para cuestionar los resultados del algoritmo, rara vez lo hacen, pues depositan su confianza en el sistema y aceptan casi

automáticamente sus predicciones, permitiendo que los sesgos algorítmicos persistan en la toma de decisiones (Eubanks, 2017).

En tercer lugar, existen los sesgos de confirmación, que ocurren cuando los funcionarios públicos tienden a aceptar las decisiones algorítmicas que refuerzan sus propias creencias o prejuicios preexistentes; en lugar de desafiar los resultados del sistema automatizado, pueden usarlos como justificación para sostener decisiones sesgadas, lo que no solo impide la corrección de los sesgos algorítmicos, sino que además contribuye a la amplificación de desigualdades estructurales (Sued, 2022). Este fenómeno se ejemplifica con el caso del COMPAS, donde los jueces y operadores del sistema penal ya tomaban decisiones influenciadas por percepciones y prejuicios raciales sobre la reincidencia delictiva; por lo tanto, el sesgo algorítmico no solo reforzó estas creencias, sino que también las legitimó bajo una apariencia de objetividad tecnológica. Como resultado, la mayoría de las decisiones generadas por el algoritmo son aceptadas sin un análisis crítico, contribuyendo así a perpetuar y profundizar la discriminación en el sistema de justicia penal de Estados Unidos (Angwin *et al*., 2016).

Además de la necesidad de que las decisiones algorítmicas sean validadas por intervención humana, se plantea la importancia fundamental de implementar auditorías algorítmicas como un mecanismo para identificar, analizar y mitigar los sesgos. Uno de los principales obstáculos para la implementación efectiva de auditorías algorítmicas es la falta de transparencia en el funcionamiento de la IA, ya que en muchos casos los algoritmos utilizados en la administración pública son desarrollados por empresas privadas cuyos modelos y códigos fuente están protegidos bajo derechos de propiedad intelectual o industrial. A esto se le conoce como opacidad algorítmica e impide que expertos independientes, reguladores y la sociedad civil puedan acceder, evaluar y cuestionar la forma en que estos sistemas toman decisiones que afectan a millones de personas. Esta situación se ejemplifica claramente en el caso del COMPAS, pues la empresa Northpointe, creadora del algoritmo, se ha negado a dar a conocer el código fuente, argumentando que está protegido por leyes de secreto comercial, impidiendo con ello que se conozca su funcionamiento y dificultando las discusiones legales en torno a su implementación.



Debido a las limitaciones de las estrategias para reducir y mitigar los sesgos algorítmicos en los modelos de predicción de riesgos utilizados en la administración pública, y ante la falta de transparencia que dificulta la realización de auditorías algorítmicas, el rol de la sociedad civil y sus organizaciones es crucial. Los casos analizados en esta investigación muestran que la participación de la sociedad civil no solamente permite evidenciar estos sesgos y sus consecuencias, sino también prevenir su reproducción y promover el desarrollo de sistemas más justos y equitativos. En el caso de sistema COMPAS, el sesgo algorítmico hacia personas afroamericanas fue evidenciado por una investigación de ProPublica, una agencia de noticias independiente y sin fines de lucro. Los sesgos hacia personas hispanas y por razones de género fueron evidenciados por investigaciones académicas independientes, mientras que el American Law Institute ha abogado por la aplicación

de normas diferenciadas para hombres y mujeres con el objetivo de neutralizar los efectos del sesgo de género. En el caso del SyRI, los sesgos fueron evidenciados por un relator especial de la ONU, el cual remitió su informe al Tribunal de La Haya, donde se declaró ilegal el uso del algoritmo. En el caso del AFST, los sesgos han sido evidenciados por investigaciones académicas independientes, lo que ha llevado a varios procesos de reformulación del modelo. Por último, los sesgos en la Plataforma Tecnológica de Intervención Social fueron evidenciados por estudios de la World Web Foundation y del Laboratorio de Inteligencia Artificial Aplicada de la UBA, los cuales fueron retomados por periodistas, organizaciones de la sociedad civil y colectivos feministas, resultando en la suspensión indefinida del algoritmo para predecir el embarazo en adolescentes.

## Conclusiones

En esta investigación se analizó una serie de casos representativos a nivel internacional para identificar y analizar las causas de los sesgos algorítmicos en los modelos de predicción de riesgo utilizados en la administración pública, así como sus principales consecuencias e implicaciones éticas, políticas y sociales. Estos sesgos pueden originarse en el proceso de diseño del algoritmo cuando existen errores en el tratamiento de las variables predictivas de riesgo o cuando se seleccionan ciertas variables como predictoras en función de los perjuicios y estereotipos presentes en los diseñadores del algoritmo. Estos sesgos también se pueden originar en la etapa de entrenamiento del algoritmo cuando se utilizan datos que reflejan perjuicios estructurales y sesgos históricos, o cuando existen problemas de representatividad en dichos datos. Los sesgos algorítmicos ocasionan la sobreestimación del riesgo de grupos sociales históricamente vulnerables, reforzando así las desigualdades y los patrones discriminatorios preexistentes en la sociedad, y consolidando dinámicas de exclusión y marginación que afectan el acceso a derechos fundamentales, programas sociales y servicios públicos esenciales. Cuando los sesgos algorítmicos se filtran en los procesos de toma de decisiones de la administración pública, el problema trasciende lo meramente técnico y se convierte en un desafío ético, político y de justicia social.



Los algoritmos de IA siempre estarán influenciados por los valores y prejuicios de sus creadores y los datos utilizados para su entrenamiento, lo que hace imposible su total neutralidad. Cada decisión en su desarrollo implica subjetividad, afectando qué información se prioriza y cómo se la procesa; esto genera construcciones estadísticas que reflejan sesgos humanos en distintos grados. Los sesgos algorítmicos no se pueden evitar, tan sólo mitigar y regular. Los casos analizados revelan que las estrategias implementadas por las distintas dependencias públicas para mitigar los sesgos algorítmicos resultaron insuficientes. Específicamente, la estrategia implementada para que los funcionarios públicos supervisen y validen las decisiones tomadas por los algoritmos, conocida como *human-in-the-loop*, se mostró ineficaz para mitigar los sesgos observados. Esto se debió a la persistencia

de diversas problemáticas relacionadas con la compleja relación que se establece entre los humanos y los sistemas automatizados, tal y como los sesgos cognitivos, los sesgos de automatización y los de confirmación, los cuales limitan la capacidad de intervención humana para corregir las distorsiones en los resultados algorítmicos.

Ante las fallas en las estrategias de mitigación, los casos analizados demuestran que la participación proactiva de la sociedad civil es fundamental para identificar y visibilizar los sesgos algorítmicos presentes en la administración pública, así como comprender sus implicaciones sociales. Además, su involucramiento no solamente contribuye a prevenir la reproducción de estas desigualdades, sino que también impulsa la creación de mecanismos de control y supervisión que fomenten el desarrollo de sistemas más justos, transparentes y equitativos. La sociedad civil se convierte en un actor clave en la construcción de una IA más ética y alineada con los principios de equidad e inclusión.

Para construir estrategias para la regulación de los sesgos algorítmicos en la administración pública, es fundamental que la implementación de estos sistemas vaya acompañada de mecanismos de transparencia, auditoría y rendición de cuentas, así como de una participación de la sociedad civil y expertos en ética digital, para garantizar que la IA en la administración pública sea una herramienta para la equidad y no un vehículo de exclusión. Diversos países han adoptado enfoques innovadores para garantizar estas condiciones. Por ejemplo, en el año 2020, Uruguay puso en marcha la Estrategia de Inteligencia Artificial para el Gobierno Digital con el propósito de impulsar y consolidar el uso responsable de la IA en la administración pública. Esta estrategia subraya la importancia de que los algoritmos implementados en el sector público respeten los derechos humanos, las libertades individuales y la diversidad. Además, adopta un enfoque de transparencia activa, lo que implica que tanto los algoritmos como los datos utilizados para su entrenamiento, junto con las pruebas y validaciones realizadas, deben estar disponibles para el público, permitiendo así auditorías algorítmicas efectivas. También se establece que cada sistema de IA debe contar con una persona claramente identificada, que asumirá la responsabilidad de sus consecuencias. Por último, la estrategia enfatiza que cualquier dilema ético derivado del uso de IA debe ser analizado y resuelto exclusivamente por seres humanos, garantizando así una toma de decisiones alineada con valores éticos y sociales.



**Reconocimiento de uso de IA**

El autor declara que, una vez concluida la redacción inicial del artículo, se utilizó la herramienta ChatGPT para resumir la información recopilada y analizada sobre los casos de estudio, cuyo desarrollo resultó particularmente extenso. El uso de esta herramienta tuvo como único propósito ajustar la extensión del trabajo al límite máximo de palabras establecido por la normativa editorial de la revista, sin que ello implicara cambios sustantivos al contenido, el análisis crítico realizado, las conclusiones y el uso de fuentes, de las que el autor se hace íntegramente responsable.

**Bibliografía**

Allhutter, D., Cech, F., Fischer, F., Grill, G. & Mager, A. (2020). Algorithmic profiling of job seekers in Austria: How austerity politics are made effective. Frontiers in Big Data, 3, 502780. DOI: https://doi.org/10.3389/fdata.2020.00005.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. ProPublica, 23 de mayo. Recuperado de: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aróstegui, L. A. (2023). El uso de RISCANVI en la toma de decisiones penitenciarias. Estudios Penales y Criminológicos, 44(Ext.), 1-43. DOI: https://doi.org/10.15304/epc.44.8884.

Arts, J. & Van den Berg, M. (2024). What the Dutch benefits scandal and policy's focus on 'fraud' can teach us about the endurance of empire. Critical Social Policy, 45(1), 177-187. DOI: https://doi.org/10.1177/02610183241281346.

Avella, M. D. P. R., Sanabria-Moyano, J. E. & Dinas-Hurtado, K. (2022). Uso del algoritmo COMPAS en el proceso penal y los riesgos a los derechos humanos. Revista Brasileira de Direito Processual Penal, 8(1), 275-310. DOI: https://doi.org/10.22197/rbdpp.v8i1.615.



Baker, R. & Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32, 1052-1092. DOI: https://doi.org/10.1007/s40593-021-00285-9.

Berning, P. A. D. (2023). El uso de sistemas basados en inteligencia artificial por las Administraciones públicas: estado actual de la cuestión y algunas propuestas ad futurum para un uso responsable. Revista de Estudios de la Administración Local y Autonómica, (20), 165-185. DOI: https://doi.org/10.24965/reala.11247.

Blanchard, M. (2022). Predictive Analytics in Child Welfare: Five Principles for Regulating Algorithmic Accountability in a New Wave of Predictive Models. University of Baltimore Law Review, 51(3), 421-448. Recuperado de: https://scholarworks.law.ubalt.edu/ublr/vol51/iss3/5.

Blomberg, T., Bales, W., Mann, K., Meldrum. R. & Nedelec Bales. J. (2010). Validation of the COMPAS risk assessment classification instrument. Tallahassee: Florida State University. Recuperado de: http://criminology.fsu.edu/wp-content/uploads/Validation-of-the-COMPASRisk-Assessment-Classification-Instrument.pdf.

CAF (2021). EXPERIENCIA. Datos e Inteligencia Artificial en el sector público. Caracas: CAF. Recuperado de: https://scioteca.caf.com/handle/123456789/1793.

Chouldechova, A., Benavides-Prado, D., Fialko, O. & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. En Conference on Fairness, Accountability and Transparency (134-148). Recuperado de: https://proceedings.mlr.press/v81/chouldechova18a.html.

Eubanks, V (2017). Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor. Nueva York: St. Martin's Press.

Feathers, T. (2023). False Alarm: How Wisconsin Uses Race and Income to Label Students "High Risk". The Markup, 27 de abril. Recuperado de: https://themarkup.org/machine-learning/2023/04/27/false-alarm-how-wisconsin-uses-race-and-income-to-label-students-high-risk.

Fuchs, D. J. (2018). The dangers of human-like bias in machine-learning algorithms. Missouri S&T's Peer to Peer, 2(1), 1-14. Recuperado de: https://scholarsmine.mst.edu/peer2peer/vol2/iss1/1.

Gerchick, M., Jegede, T., Shah, T., Gutierrez, A., Beiers, S., Shemtov, N., Xu, K., Samant, A. & Horowitz, A. (2023). The devil is in the details: Interrogating values embedded in the allegheny family screening tool. En Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (1292-1310). DOI: https://doi.org/10.1145/3593013.3594081.

Hagerty, A., Aranda, F. & Jemio, D. (2023). Predictive puericulture in Argentina: The Plataforma Tecnológica de Intervención Social and the reproduction of Latin eugenics. BJHS Themes, 8, 221-235. DOI: https://doi.org/10.1017/bjt.2023.2.



Hamilton, M. (2019). The sexist algorithm. Behavioral sciences & the law, 37(2), 145-157. DOI: https://doi.org/10.1002/bsl.2406.

Hamilton, M. (2019). The biased algorithm: Evidence of disparate impact on Hispanics. American Criminal Law Review, 56(4), 1553-1577. Recuperado de: https://www.law.georgetown.edu/american-criminal-law-review/wp-content/uploads/sites/15/2019/06/56-4-the-biased-algorithm-evidence-of-disparate-impact-on-hispanics.pdf.

Johansen, J., Pedersen, T. & Johansen, C. (2020). Studying the transfer of biases from programmers to programs. ArXiv preprint arXiv:2005.08231. DOI: https://doi.org/10.48550/arXiv.2005.08231.

Koniakou, V. (2023). From the "rush to ethics" to the "race for governance" in Artificial Intelligence. Information Systems Frontiers, 25(1), 71-102. DOI: https://doi.org/10.1007/s10796-022-10300-6.

Kordzadeh, N. & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388-409. DOI: https://doi.org/10.1080/0960085X.2021.1927212.

Lappalainen, K. F. (2021). Protecting Children from Maltreatment with the Help of Artificial Intelligence: A Promise or a Threat to Children's Rights? En de Vries, K. & Dahlberg, M. (Eds.), De Lege. Law, AI and Digitalisation (431-466). DOI: https://doi.org/10.33063/dl.vi.433.

Lazcoz Moratinos, G., & Castillo Parrilla, J. A. (2020). Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: el caso SyRI. Revista chilena de derecho y tecnología, 9(1), 207-225. DOI: http://dx.doi.org/10.5354/0719-2584.2020.56843.

Marinucci, L., Mazzuca, C. & Gangemi, A. (2023). Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender. AI & Society, 38(2), 747-761. DOI: https://doi.org/10.1007/s00146-022-01474-3.

Martín, J. (2022). Inteligencia artificial, sesgos y no discriminación en el ámbito de la inspección tributaria. Crónica Tributaria, (182/2022), 51-89. DOI: https://doi.org/10.47092/CT.22.1.2.

Nadeem, A., Olivera, M. & Babak, A. (2022). Gender bias in AI-based decision-making systems: a systematic literature review. Australasian Journal of Information Systems, 26, 1-34. DOI: https://doi.org/10.3127/ajis.v26i0.3835.



Ortiz Freuler, J. & Iglesias, C. (2018). Algorithms and Artificial Intelligence in Latin America: A Study of Implementation by Governments in Argentina and Uruguay. Washington DC: World Wide Web Foundation. Recuperado de: https://webfoundation.org/docs/2018/09/WF_AI-in-LA_Report_Screen_AW.pdf.

Pasquinelli, M., Cafassi, E., Monti, C., Peckaitis, H. & Zarauza, G. (2022). Cómo una máquina aprende y falla. Una gramática del error para la Inteligencia Artificial. Revista Hipertextos, 10(17), 13-29. DOI: https://doi.org/10.24215/23143924e054.

Pedace, K., Schleider, T. J. & Balmaceda, T. (2023). Inteligencia artificial y sesgos. El caso de la predicción del embarazo adolescente en Salta. Revista Iberoamericana de Ciencia, Tecnología y Sociedad, 18(53), 9-26. DOI: https://doi.org/10.52712/issn.1850-0013-359.

Peeters, R. & Widlak, A. (2023). Administrative exclusion in the infrastructure- level bureaucracy: The case of the Dutch daycare benefit scandal. Public Administration Review, 83(4), 863-877. DOI: https://doi.org/10.1111/puar.13615.

Red en Defensa de los Derechos Digitales (R3D) (2020). Países Bajos suspende uso de algoritmo por discriminar grupos vulnerables. Red en Defensa de los Derechos Digitales, 17 de febrero. Recuperado de: https://r3d.mx/2020/02/17/paises-bajos-suspende-uso-de-algoritmo-por-discriminar-grupos-vulnerables/.

Rittenhouse, K., Putnam-Hornstein, E. & Vaithianathan, R. (2023). Algorithms, Humans and Racial Disparities in Child Protective Systems: Evidence from the Allegheny Family Screening Tool. Documento de trabajo.

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M. & Gregor, S. (2022). Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. European Journal of Information Systems, 31(3), 313-338. DOI: https://doi.org/10.1080/0960085X.2021.1960905.

Savage, M. (2021). DWP urged to reveal algorithm that 'targets' disabled for benefit fraud. The Guardian, 21 de noviembre. Recuperado de: https://www.theguardian.com/society/2021/nov/21/dwp-urged-to-reveal-algorithm-that-targets-disabled-for-benefit.

Simon, J., Hang Wong, P. & Rieder. G. (2020). Algorithmic bias and the Value Sensitive Design approach. Internet Policy Review, 9(4), 1-16. DOI: https://doi.org/10.14763/2020.4.1534.

Skeem, J., Monahan, J. & Lowenkamp, C. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. Law and Human Behavior, 40(5), 580-593. DOI: https://doi.org/10.1037/lhb0000206.

Sued, G. (2022). Culturas algorítmicas: conceptos y métodos para su estudio social. Revista Mexicana de Ciencias Políticas y Sociales, 67(246), 43-73. DOI: https://doi.org/10.22201/fcpys.2448492xe.2022.246.78422.



Valle-Cruz, D., García-Contreras, R. & Gil-Garcia, J. R. (2024). Exploring the negative impacts of artificial intelligence in government: the dark side of intelligent algorithms and cognitive machines. International Review of Administrative Sciences, 90(2), 353-368. DOI: https://doi.org/10.1177/0020852323118705.

Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. Colorado Technology Law Journal, 17, 131-160. Recuperado de: https://ctlj.colorado.edu/wp-content/uploads/2021/02/17.1_4-Washington_3.18.19.pdf.