Sesgo de género en la inteligencia artificial: "de tal palo, tal astilla"*

Gender bias in artificial intelligence: "from such a stick, such a splinter"

Melanie Muriel Freres Hellebaut**

Resumen

La inteligencia artificial fue concebida por el ser humano para facilitar tareas de procesamiento y clasificación de datos y dar soluciones rápidas a determinados problemas. No obstante, como su creador, es imperfecta y tiende a repetir patrones que pueden ser discriminatorios. Esta falla en el sistema se debe tanto a su construcción como a la información que la alimenta y, hasta nuestros días, no ha podido ser superada, pese a las diversas soluciones que se han planteado al respecto.

Palabras clave: Sesgo, género, inteligencia artificial, dato, algoritmo

Cómo citar este artículo: Freres, M. M. (2021). Sesgo de género en la inteligencia artificial: "de tal palo, tal astilla". *Revista Nueva Época*, 57, 95-116. https://doi.org/10.18041/0124-0013/nuevaepoca.57.2021.9107

* Este artículo ha sido desarrollado, previamente, por la autora como trabajo final de investigación en el curso sobre Derecho Civil y Protección de Datos, para el Doctorado en Derecho de la Universidad de Buenos Aires.

** La autora es abogada. Licenciada en Ciencias Jurídicas y Sociales por la Universidad de Concepción, Chile. Diplomada en Pedagogía para la Educación Superior por la Universidad Santo Tomás, La Serena, Chile. Postitulada en Gestión Resolutiva de conflictos jurídicos y magíster en Arbitraje por la Universidad Central, Santiago, Chile. Actualmente cursa para el Doctorado en Derecho de la Universidad de Buenos Aires, Argentina, es miembro del Directorio de la Asociación Gremial de Árbitros Concursales y Contractuales de Chile y profesora de Derecho Comercial en la Universidad Católica del Norte, sede Coquimbo, y en la Universidad del Alba, sede La Serena, Chile. Dirección postal: Prat N° 216, of. B, La Serena; teléfono: 56977954600. Correo electrónico: melaniefreres@gmail.com

Abstract

Artificial Intelligence was conceived by humans to facilitate data processing and classification tasks in order to provide quick solutions to certain problems. However, like its creator, it is imperfect and tends to repeat patterns that can be discriminatory. This flaw in the system is due both to its construction and to the information that feeds it and, to this day, it has not been overcome, despite the various solutions that have been proposed in this regard.

Keywords: Bias, gender, artificial intelligence, data, algorithm

1. Introducción

La Real Academia Española define *sesgo* como "una oblicuidad o torcimiento", es un "concepto que se utiliza simbólicamente para mencionar o referenciar una tendencia o una inclinación hacia algo" (Macchiavelli, 2021, p.).

Por su parte, la activista Marta Lamas (2012) ha señalado que *género* es:

> [...] el conjunto de creencias, prescripciones y atribuciones que se construyen socialmente tomando a la diferencia sexual como base. Construcción social que funciona como una especie de filtro cultural con el cual se interpreta al mundo y, también, como una especie de armadura con la que se constriñen las decisiones y oportunidades de las personas dependiendo de si tienen cuerpo de mujer o cuerpo de hombre. (p.)

En otra de sus obras, la autora agrega que "el género produce un imaginario social con una eficacia simbólica contundente que, al dar lugar a concepciones sociales y culturales sobre la masculinidad y feminidad, es usado para justificar la discriminación por sexo y por prácticas sexuales" (Lamas, 2000, p.).

Para la conmemoración del 25º aniversario de la adopción de la Declaración y Plataforma de Acción de Beijing, el Programa de las Naciones Unidas para el Desarrollo publicó un exhaustivo informe sobre el estado de la desigualdad de género en el mundo, mediante el cual se concluye que, "a pesar de los notables progresos realizados en algunas esferas, los sesgos y los prejuicios contra las mujeres persisten e, incluso, han empeorado en los últimos años" (Organización de las Naciones Unidas —ONU—, 2020, p.)[1].

[1] Traducción libre.

Los resultados son abrumadores: el 91 % de los hombres y el 86 % de las mujeres siguen teniendo algún tipo de sesgo discriminatorio hacia las mujeres, los que difieren en su contenido generalmente político, educacional o económico. Según el índice, aproximadamente la mitad de los hombres y mujeres de los 75 países escrutados —que representan más del 80 % de la población mundial— piensan que los hombres son mejores líderes políticos. Más del 40 % siente que los hombres son mejores ejecutivos de negocios y que tienen más derecho al trabajo cuando los empleos son escasos. Increíblemente, el 28 % justifica que un hombre golpee a su mujer (Organización de las Naciones Unidas —ONU—, 2020).

Bajo aquel paradigma, el problema científico planteado en esta investigación es si el sesgo de género del que venimos hablando se proyecta también al ámbito de la inteligencia artificial, en adelante, IA.

Nuestra hipótesis descansa sobre la respuesta afirmativa, recayendo el objetivo general del presente trabajo en determinar si efectivamente existe sesgo de género en la IA, siendo sus objetivos específicos analizar el sesgo; identificar la forma, el tiempo y el lugar en que eventualmente se incrustaría en la IA; describir las soluciones propuestas por los organismos internacionales y la doctrina e indagar acerca de la posición asumida al efecto por los tribunales de justicia a nivel comparativo.

En consecuencia, la investigación es de tipo exploratoria, pues recae sobre un problema respecto del cual, en relación con el estado del arte, existe un mínimo de ensayos previos. Es explicativa en cuanto al objeto de estudio y descriptiva en relación con las posibles soluciones planteadas e indagatoria acerca del estado de la jurisprudencia.

Por cierto, el tema en estudio se encuentra inscrito en un marco teórico mayor, como es el sesgo de género, mientras que el marco conceptual se integra con la estipulación de significado de los siguientes términos que se reputan trascendentes en el desarrollo de la investigación: inteligencia artificial, dato y algoritmo.

Para llevar a cabo el trabajo, se utilizó la metodología cualitativa, mediante la técnica de investigación documental, por medio de la observación directa de fuentes bibliográficas disponibles en internet, dadas las circunstancias impuestas por la pandemia.

Pese a la escasa producción científica disponible, en cuanto al estado de la cuestión, podríamos decir que, sin duda, constan avances en su conoci-

miento, pero sin un debate relevante en cuanto a las posiciones esgrimidas por los diversos autores que encuadran el marco teórico son muy similares.

A fin de cumplir con nuestros objetivos, haremos, primeramente, una contextualización histórico-teórica de la IA, para luego entrar al análisis de los sesgos, a la descripción de las propuestas de solución y a la determinación del estado de la jurisprudencia en relación con el objeto de estudio.

2. Desarrollo

2.1 Breve reseña histórica

Los orígenes de la IA se remontan a finales del siglo XIX, en efecto, ya en el año 1854 el matemático George Boole argumenta, por primera vez en la historia, que el razonamiento lógico podría sistematizarse de la misma manera que se resuelve un sistema de ecuaciones (National Geographic, 2019).

Más tarde, Alan Turing, considerado padre de la computación moderna, publica, en 1936, un artículo sobre los números computables en el que introduce el concepto de algoritmo y sienta las bases de la informática. Luego, en su ensayo de 1950, titulado "Computing Machinery and Intelligence", conocido como el "Test de Turing", propone una prueba de comunicación verbal hombre-máquina que evalúa la capacidad de estas últimas de hacerse pasar por humanos (National Geographic, 2019).

Entretanto, Konrad Zuse crea, en 1941, la primera computadora programable y completamente automática. Posteriormente, en 1956, el informático John McCarthy acuña inicialmente el término *inteligencia artificial* durante la conferencia de Darmouth, en Estados Unidos. Al año siguiente, Frank Rosenblat diseña la primera red neuronal artificial (National Geographic, 2019).

En 1943, Warren McCulloch y Walter Pitts propusieron uno de los primeros modelos matemáticos de una neurona, a partir del cual se han inspirado las actuales redes neuronales artificiales (Prieto *et al.*, 2000).

2.2 La inteligencia artificial

Podemos definir a la IA como:

> [...] una familia de técnicas que son utilizadas para simular o recrear la inteligencia humana a través de las ciencias computacionales con la finalidad de que los sistemas brinden velocidad de procesamiento y de clasificación de datos, dándonos respuestas o propuestas rápidas. (Macchiavelli, 2021, p.)

Las técnicas de IA son variadas, entre ellas encontramos: el *machine learning* o el aprendizaje automático; el *fuzzy logic* o lógica difusa; la vida artificial; los sistemas expertos; las redes Bayesianas o de creencia; la ingeniería del conocimiento; las redes neuronales artificiales; los sistemas reactivos; los sistemas basados en reglas; el razonamiento basado en casos; la lingüística computacional y el procesamiento del lenguaje natural (Asociación para el Progreso de la Dirección. España, 2021).

El *machine learning* es:

> [...] la rama de la ciencia que busca el desarrollo de técnicas de IA que permitan a los ordenadores aprender por sí mismos. Para ello se crean programas que pueden generalizar ciertas respuestas a partir de información sin estructurar, que se suministra como ejemplos, así se induce al conocimiento por parte del ordenador. (Asociación para el Progreso de la Dirección. España, 2021, p.)

Este aprendizaje:

> [...] puede ser supervisado por la propia máquina, trabajando con datos de los que aprende a asignar una etiqueta de salida. Si, por el contrario, no dispone de datos etiquetados, la máquina infiere del conjunto de datos mediante asociaciones de manera no supervisada. Si se entrena al algoritmo a partir de la prueba y el error, se le llama aprendizaje por refuerzo. (Macchiavelli, 2021, p.)

Citando a García Moreno, Macchiavelli (2021) agrega que:

> [...] el aprendizaje automático tiene como variante el *Deep Learning* o aprendizaje profundo, que es una forma de entrenamiento sofisticado en que la máquina aprende usando una red neuronal artificial, asignando un número de niveles jerárquicos. En el nivel más bajo, la red aprende algo simple y envía esa información al siguiente nivel en el que tomará esa información y la combinará obteniendo una información más compleja y así sucesivamente hasta llegar al nivel más elevado. (p.)

Continúa diciendo que este:

> [...] aprendizaje profundo recibe el nombre de "caja negra" cuando es incapaz de explicar la secuencia o los pasos del razonamiento que ha seguido para llegar a la respuesta ofrecida, lo que no logra hacer porque la ruta que recorren los algoritmos para tomar una decisión incluye parámetros y operaciones complejas que dificultan su comprensión. (Macchiavelli, 2021, p.)

Sin embargo, "existen técnicas inteligentes tales como la automatización o la detección inteligente que, a partir del uso de las denominadas "cajas blancas", sí permiten explicar y trazar el modo de razonamiento" (Almagro Blanco, 2018, p.).

Así las cosas,

> [...] con el uso de técnicas de cajas negras, donde los algoritmos se retroalimentan y aprenden solos, desarrollando soluciones que pueden ser discriminatorias, podemos llegar a un problema de sesgos. Si bien las técnicas de caja blanca, permiten tener más controlados los resultados, puesto que los algoritmos razonan aquello que previamente se les enseñó, pudiendo supervisarse e intervenir para una respuesta correcta, esta s también pueden presentar sesgos, pero, al menos, no será producto de los patrones asociados por los algoritmos sin la intervención humana. (Macchiavelli, 2021, p.)

Lo anterior demuestra que la IA no es verdaderamente objetiva, toda vez que los algoritmos y datos que la conforman pueden reflejar los sesgos de quienes los crean y proporcionan, "incluso aquellos que son imparciales al principio, pueden aprender los sesgos de sus entrenadores humanos a lo largo del tiempo" (Rivas y Espinoza, 2020, p.).

Y esto por cuanto, para crear IA, se requieren dos elementos básicos: los algoritmos y los datos para configurarlos. El algoritmo proporciona las instrucciones para la máquina y los datos permiten a la máquina aprender a emplear esas instrucciones y perfeccionar su uso (KeepCoding, 2022). De esta forma,

> [...] el algoritmo como procedimiento sistemático y efectivo para resolver un problema se compone de un número concreto de instrucciones que deben estar bien definidas para operar sobre un tipo concreto de datos a través de un número finito de pasos que aporta una solución, generalmente matemática, a todos los casos analizados. (Benítez, 2019, p.)

En palabras simples, un algoritmo se conforma por "un conjunto de instrucciones informáticas que recibe una máquina para realizar una acción o resolver un problema, por medio de la deducción, la búsqueda, la clasificación o la comunicación" (Cognodata, 2019, p.). Por lo tanto,

> [...] un elemento esencial para este subcampo de la IA son los datos, que constituyen la materia prima utilizada para automatizar el proceso de aprendizaje en el que los sistemas son entrenados para realizar predicciones, estos pue-

> den ser imágenes, sonidos, texto escrito, redes, posiciones de un *GPS*, tablas o cualquier otro tipo de representación. (Ferrante, 2021, p.)

De tal manera que los sistemas de IA pueden replicar sesgos sociales existentes, con el potencial de exacerbarlos, pues la mayoría de ellos se construyen usando conjuntos de datos que reflejan el mundo real, que suele ser imperfecto, injusto y discriminatorio. Hecho que es peligroso, no solo porque ellos perpetuarían esas desigualdades, sino por incrustarlas en formas opacas, en cuanto que el sistema es falsamente aclamado como "objetivo" y "preciso" (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura —UNESCO—, 2019a).

Estas desigualdades son, principalmente, el resultado de cómo aprenden las máquinas. De hecho, como el aprendizaje automático se basa en los datos que las alimentan, se necesita especial atención para promover y recopilar datos sensibles, prestando atención a los datos sesgados a fin de limitar los puntos ciegos (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura —UNESCO—, 2019a).

Consecuencialmente, las desigualdades de género comienzan tempranamente en la conceptualización y el diseño de sistemas de IA. Esto se debe, en gran medida, a que las mujeres siguen siendo minoría en la fuerza laboral de las mismas, así como en la generación de datos y algoritmos y en su auditoría y control, por lo que sus voces no son representadas equitativamente en los procesos de toma de decisiones que intervienen en el diseño y desarrollo de sistemas de IA, con el consiguiente riesgo de elaborar tecnologías solo para algunos datos demográficos. Además, los sesgos que las personas llevan en su vida cotidiana pueden también reflejarse e incluso amplificarse a través del desarrollo y el uso de sistemas de IA (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura —UNESCO—, 2019a).

Hay algo que es aún peor: la noción de *género*, en los sistemas de IA, suele ser una elección binaria, lo que implica necesariamente ignorar y excluir activamente a las personas LGTBIQ+, llegando a discriminarlos en formas humillantes (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura —UNESCO—, 2019a)[2]. Cabe subrayar que, hoy en día, el género no puede ser delimitado de manera taxativa. Con "dato de género" nos podemos estar refiriendo al sexo biológico, a la identidad de género, a la orientación sexual, al rol

[2] Traducción libre.

de género o a la expresión de género o a cualquier procesamiento que pueda revelar estas categorías de forma directa o indirecta tiene un potencial efecto discriminatorio (Faliero, 2021).

2.3 El sesgo en la inteligencia artificial

Por su parte, el "sesgo algorítmico":

> [...] ocurre cuando un sistema informático refleja los valores o —disvalores— de los humanos que están implicados en la codificación y recolección de datos usados, para entrenar el algoritmo. Estos se encuentran en todas partes, por lo que pueden ocasionar un gran impacto en temas, como la privacidad o agravar sesgos sociales existentes respecto a razas, género, sexualidad o etnias.
>
> Los problemas de comprensión, investigación y descubrimiento de sesgos en los algoritmos provienen de la naturaleza de estos, ya que, sus propietarios no suelen permitir el acceso a su implementación y muchas veces son demasiado complejos para entender cómo funcionan. Además, los algoritmos pueden cambiar o responder a diferentes entradas de maneras imprevistas o fácilmente reproducidas para su análisis. En muchos casos, dentro de un solo sitio *web* o aplicación no hay un solo algoritmo para examinar, sino todo un conjunto de procesos y entradas de datos interrelacionados. ("Sesgo algorítmico", 2022, p.)

Ahora bien, el sesgo algorítmico se relaciona directamente con:

> [el] proceso humano neurobiológico de toma de decisiones, por medio del cual, los sistemas terminan por aprender de nuestras preferencias, replicándolas. Si nuestras decisiones prefieren o se inclinan por discriminar a personas, eso se verá plasmado en datos o en conductas que serán tomados por los sistemas informáticos para proponer o tomar decisiones. Consiguientemente, el problema de la discriminación digital que proviene de los sesgos algorítmicos se origina, precisamente, en la toma de decisiones humanas. (Macchiavelli, 2021, p.)
>
> Desde la mirada de las neurociencias, la evidencia científica demuestra que las personas decidimos, básicamente, con las emociones antes que con la razón, por cuanto, la toma de decisiones es un mecanismo complejo en el que la emoción facilita el proceso, en el que influyen el estado de ánimo y el déficit en la función cognitiva. (Manes y Niro, 2014, p.)

> Ello sucede porque en el proceso de toma de decisiones, se activan áreas cerebrales involucradas en el control de las emociones y que, además, por la velocidad en que esto ocurre, no cabría tiempo para racionalizar", por lo que "se producirían procesos implícitos inconscientes en los que el cerebro toma decisiones incluso antes de que seamos conscientes. (Manes y Niro, 2014, p.)

A lo anterior cabe agregar "la importancia del factor contexto en la toma de decisiones" (Macchiavelli, 2021, p.), pues "los seres humanos, basados en nuestra experiencia, intuición, aprendizaje y emoción, integramos la información en un contexto que cambia permanentemente de manera inmediata y automática" (Manes y Niro, 2014, p.).

Citando a Martín Alfonso, Macchiavelli (2021) señala que

> [...] la psicología explica que, cuando las personas socializamos aprendemos reglas para intentar dar sentido a lo que nos rodea, las que nos sirven de filtro para procesar la información. Si esas reglas están distorsionadas, serán fuente de sesgos que influirán en la interpretación de la realidad como son las etiquetas o las generalizaciones extremas. (p.)

> Si bien existe una extensa lista sobre los diferentes tipos y subtipos de sesgos, el estadístico, el cultural y el cognitivo son los que mejor sirven para explicar por qué se producen las inclinaciones o las tendencias en el proceso de toma de decisiones y, particularmente, su impacto en la tecnología en cuestiones de género. (Macchiavelli, 2021, p.)

"El sesgo estadístico que intenta recrear la imagen, dice estrecha relación con el cómo se obtienen datos" (Macchiavelli, 2021, p.). "Conviene recordar en este punto la importancia de reconocer las diferentes etapas en las que el sesgo en tecnología logra aparecer" (Hao, 2019, p.), "sea en la de identificación del objetivo cuyo fin puede estar sesgado, como también en la recogida o preparación de datos que le han servido de ancla" (Macchiavelli, 2021, p.).

En estos casos, vale decirse que:

> [...] se incurre en un error sistémico donde se facilita la inclinación hacia un resultado determinado" que, por lo general, no constituye un error casual en lo que a cuestiones de género respecta, desde que se encuentra en la recogida de información o datos, lo que se halla muy "relacionado con los

denominados sesgos culturales y cognitivos, puesto que estos también contribuyen al impacto de las diferentes formas de discriminación. (Macchiavelli, 2021, p.)

En efecto, el sesgo cultural es aquel que deriva de la sociedad, del lenguaje o de todo lo que hemos aprendido a lo largo de nuestras vidas. Los estereotipos de género, son un ejemplo de ello. Históricamente, lo doméstico y lo servicial ha sido atribuido a roles asociados al género, en tanto se daba por hecho que esas eran tareas propias de mujeres, lo cual refuerza la idea de que el problema no es solo de la tecnología ni de las técnicas en sí, sino que se asocia a un problema cultural que, en estos casos, termina por impactar en la tecnología. (Macchiavelli, 2021, p.)

"Por último, el sesgo cognitivo es asociado a nuestras preferencias. Se trata de esquemas mentales que ayudan al cerebro a procesar la información y dar respuesta a situaciones a las que se debe enfrentar de manera rápida" (Manes, 2020, p.).

El problema de estos esquemas es que percibimos la realidad con nuestros filtros culturales aun frente a datos objetivos que nos indiquen que estamos equivocados, por lo que nuestra mente intentará que nuestras ideas personales encajen de algún modo para justificarnos. (p.)

Esto es lo que se conoce como "sesgo de confirmación" (Macchiavelli, 2021, p.).

2.4 En la búsqueda de soluciones

En caso de detectar posibles sesgos o rendimientos dispares en los sistemas de IA, es posible pensar en soluciones para mitigarlos. Una de ellas sería "balancear de alguna forma los datos, para evitar que los modelos resulten discriminatorios o injustos" (Ferrante, 2021, p.). Otra opción podría ser "inducir al sistema a que utilice representaciones "justas" de los datos, en el sentido de que no estén asociadas a las características que son fuente de discriminación" (Ferrante, 2021, p.).

Una tercera alternativa sería:

> [...] obligarlo directamente a ignorar los atributos protegidos, como el género u otras características demográficas, al momento de tomar una decisión. Sin embargo, aunque ocultemos ciertos atributos a un sistema, como el género o el grupo étnico al que pertenece una persona, la correlación entre estos y

> otras variables seguirá existiendo. (Ferrante, 2021, p.)

Otros sostienen que:

> [...] para superar el problema de los sesgos humanos y de los algoritmos, a más del uso responsable de la IA para mitigar los sesgos y evitar con ello las desigualdades o injusticias, se puede utilizar la propia tecnología, creando algoritmos que corrijan dichos errores. (Macchiavelli, 2021, p.)

> Lo que se conoce como *Fairness Contrains* o equidad algorítmica, intenta minimizar el riesgo de discriminación de los clasificadores de datos para que, las variables sensibles, no influyan de manera injusta en el resultado. En otros casos se propone crear modelos que reconozcan la discriminación a partir de clasificar reglas de asociación para el análisis de datos imparciales y limpiar un conjunto de datos sesgados de entrenamiento o agregar la variable de la etiqueta imparcial al modelo probabilístico bayesiano. (Macchiavelli, 2021, p.)

> Recordemos que si hay algo que los modelos de aprendizaje automático hacen bien es encontrar patrones y también correlaciones. Por eso, aun cuando la comunidad académica de investigación en equidad algorítmica *Fairness* ha trabajado arduamente durante los últimos años en la construcción de modelos justos y que no discriminen, el factor humano en el diseño de estos sistemas resulta primordial. (Ferrante, 2021, p.)

Empero, "aunque en la actualidad existen diversas formalizaciones del concepto de *Fairness*, muchas de ellas resultan mutuamente incompatibles, en el sentido de que no es posible utilizarlas al mismo tiempo" (Friedler *et al*., 2021, p.) y, por tanto, se debe optar por aquellas que se desee maximizar" (Ferrante, 2021).

No resulta suficiente

> [...] generar bases de datos representativas o modelos justos en algún sentido específico. Los sistemas de IA están diseñados por personas con sus propias visiones del mundo, prejuicios, valoraciones de los hechos y sesgos adquiridos a lo largo de su experiencia de vida, que pueden filtrarse en el diseño y la definición de criterios de evaluación para estos modelos. Si esos grupos de trabajo no son lo suficientemente diversos como para reflejar una amplia variedad de visiones, muy probablemente no lleguen siquiera a darse cuenta de la existencia de los sesgos y, por tanto, a corregirlos. (Ferrante, 2021, p.)

Ahora bien,

> [...] si la diversidad en los equipos que conciben estos sistemas resulta tan relevante, esperaríamos que en la práctica esos grupos fueran realmente heterogéneos, no solo en términos de género, sino también de clases sociales, etnias, creencias, edad u orientación sexual, solo por dar algunos ejemplos. (Ferrante, 2021, p.)

Cuestión que no se ve reflejada, actualmente, en las estadísticas.

De todas formas,

> [...] los esfuerzos realizados en los últimos años para crear conciencia sobre estos riesgos y aumentar la diversidad de la comunidad de IA, tanto en el ámbito académico como en la industria, comienzan a sentar las bases para un futuro más promisorio. Los gobiernos, por su parte, han empezado a preocuparse por la necesidad de regular el uso y desarrollo de estas tecnologías. La emergencia de foros de discusión especializados en estas temáticas y el interés de todas las ramas de la ciencia por conocer las implicancias y potenciales aplicaciones de la IA en sus propios campos de estudio han abierto nuevos horizontes para el desarrollo científico guiado por los datos. (Ferrante, 2021, p.)

"El estudio del sesgo algorítmico se encuentra enfocado especialmente en aquellos algoritmos que reflejan "discriminación sistemática e injusta". Este tipo de sesgos han empezado a ser tenidos en cuenta en marcos legales" ("Sesgo algorítmico", 2022, p.), como el Reglamento General de Protección de Datos, conocido como *RGPD*, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos del año 2016.

Por su parte, la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura recomienda medidas multidisciplinarias, concibiendo como línea trasversal a los derechos humanos de las personas. Para ello, destaca la necesidad de implementar regulaciones en el sector público, en el privado y en la comunidad tecnológica, además de mecanismos de transparencia y de monitoreo constantes de la aplicación de la IA y los algoritmos, en todos los sectores, a fin de combatir los sesgos y actos de discriminación que pudieran derivarse de una mala introducción de estas tecnologías.

Se estima de vital importancia que los medios de comunicación fomenten una mayor capacitación de los periodistas en este campo, de manera que informen con más herramientas sobre su potencial y sus amenazas para formar

ciudadanos con un rol más activo en la vigilancia de estas tecnologías y en el respeto a los derechos humanos (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura —UNESCO—, 2019b).

En el mes de febrero de 2020, la Unión Europea abordó, por medio del "Libro Blanco sobre la inteligencia artificial", el impacto que tendrá esta tecnología tanto a nivel europeo como mundial. Con una serie de propuestas que hacen énfasis en el carácter complejo y opaco de gran parte de las plataformas de IA, el organismo demanda la necesidad de un registro continuo de los algoritmos y los datos usados para el entrenamiento de sus programas, a objeto de tener la posibilidad de examinar las metodologías de programación usadas para construir estos sistemas y así evitar sesgos que deriven en actos discriminatorios (Comisión Europea, 2020a).

Un mes más tarde, en la comunicación que hace la Comisión Europea al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones, se manifiesta la preocupación acerca de que los sesgos algorítmicos propician un acceso desigual de las mujeres a los derechos humanos en comparación con sus pares masculinos. Es así como se manifiesta la necesidad de garantizar sistemas de aprendizaje automático y algoritmos de IA robustos y transparentes para evitar la existencia de prejuicios subyacentes en estas tecnologías (Comisión Europea, 2020b).

Desde la doctrina, también se ofrecen posibles soluciones a este defecto, comenzando por que el propio ser humano sea capaz de superar los estereotipos, los tratos discriminatorios y romper el sesgo de género. Esto supone una perspectiva de derechos humanos para la IA, toda vez que estos permitirían establecer un primer punto de partida muy importante en la reflexión y la resolución del problema (Rivas y Espinoza, 2020).

En segundo término, no debe subestimarse el sesgo algorítmico habida especial consideración de que estas tecnologías se aplican sobre grandes volúmenes de datos y respecto de millones de ciudadanos, provocando un efecto expansivo y multiplicador de daños que deben ser reparados (Faliero, 2021).

El tratamiento analítico de nuestros datos genera situaciones potencialmente discriminatorias que deben ser visibilizadas y denunciadas. Debe aplicarse un mayor control y una mayor auditoría de los medios digitales conjuntamente con más y mejores medios de protección. Una vez identificadas las fallas en los modelos predictivos, estas han de ser erradicadas o bien

no utilizar a estos últimos en áreas críticas, donde pueden ser vulnerados derechos humanos fundamentales, lo mismo si sus impactos son negativos (Faliero, 2021).

Se hace necesario, además, implantar algunos cambios en esta materia, tales como la reestructuración de la protección de datos y la redefinición de dato sensible, además de la implementación del derecho al anonimato, a no ser perfilado, a no ser objeto de decisiones automatizadas, a la autodeterminación informativa dinámica y al consentimiento informado dinámico. Esto, ya no como meras recomendaciones éticas, sino como imperativos normativos reconocidos en los ordenamientos jurídicos. En ese sentido, el principio de la no discriminación debiera incluir de manera expresa la prohibición de discriminación algorítmica de género (Faliero, 2021).

Se suman a las propuestas anteriores la intervención humana, con perspectiva de género, tanto para validar los datos que nutren al sistema como para revisar y contextualizar las soluciones que ofrecen los algoritmos, que pueden presentar fallas o "reproducir los sesgos humanos que dan paso a la discriminación digital" (Macchiavelli, 2021, p.). Esto, mediante la incorporación, no solo de más mujeres y expertas en tecnología, sino de personas de diversos saberes y múltiples disciplinas, idealmente con especialización en cuestiones de género y diversidad (Macchiavelli, 2021).

Otros recomiendan también:

> [...] la comprobación y eliminación de los sesgos en la compra de datos, así como la exigencia en los contratos con proveedores de cláusulas contra los sesgos; la adopción de técnicas que eviten la discriminación, potenciando un sistema de clasificación de la información legítima para dar mayor calidad a los datos; la transparencia en las políticas de privacidad y en el consentimiento de los usuarios sobre el uso de datos; el testeo de los algoritmos contra los sesgos, con "desaprendizaje" de relaciones causales y de serendipia que refuerzan la presencia de comportamientos anteriores. (Benítez, 2019, p.)

> [Aparte de la] aplicación del análisis de componentes independientes o ICA, para revelar factores ocultos en las variables, que evalúen el impacto social, cultural y político de los resultados que ofrecen los algoritmos; la investigación sobre las soluciones facilitadas por algoritmos para conocer su calidad y sus impactos para adoptar decisiones de carácter ético; y, sin duda, la necesidad de implicarse,

> aprender y comprender los efectos de los algoritmos, así como de otras transformaciones de la digitalización a fin de no caer en manos de quienes los formulan y de la propia tecnología. (Benítez, 2019, p.)

En otras palabras, actuar con conocimiento y empoderamiento en los medios digitales, tomando conciencia de la importancia del autocuidado en estos ambientes en los que plasmamos gran parte de nuestras vidas (Faliero, 2021).

2.5 Estado de la jurisprudencia

Revisadas las bases de datos de las cortes supremas de diversos países, entre ellos, Estados Unidos, Argentina y Chile, no encontramos fallos en relación directa con nuestra materia de estudio. No obstante, existe una "sentencia pionera en el ámbito digital" (The Technolawgist, 2020) dictada por la Corte de Distrito de La Haya, con fecha del 5 de febrero de 2020, en causa C-09-550982-HA ES 18-388[3], que se aproxima bastante, por cuanto recae sobre el uso de un algoritmo acusado de estigmatizar y discriminar a los individuos más vulnerables de los Países Bajos.

En esta causa, el Comité Jurídico de Derechos Humanos de ese país, junto con otras asociaciones y dos ciudadanos particulares, denunciaron al Estado neerlandés por el uso del sistema algorítmico de indicación de riesgos conocido como *System Risk Indication* o "SyRI", para combatir y prevenir fraudes en materia laboral, de seguridad social y tributaria, por medio de la predicción de la probabilidad de que determinados solicitantes de beneficios estatales intentaran burlar el sistema (The Technolawgist, 2020).

Según el Gobierno, SyRI (acrónimo de System Risk Indication), amparada en el artículo 65 de la Ley SUWI, sobre la estructura de la organización de implementación del trabajo y los ingresos,

> [...] conforma una infraestructura técnica y procedimientos asociados, que permiten vincular y analizar datos de forma anónima en un entorno seguro para poder generar informes de riesgo, entendiendo como tales, la consideración de que una persona jurídica o física merecía ser investigada en relación con un posible fraude o incumplimiento de la legislación social, laboral o fiscal pertinente. (Van Kordelaar, 2020, p.)

Por su parte, los demandantes sostienen que SyRI es incompatible con

[3] El texto completo de la sentencia se encuentra en la página web del Poder Judicial de los Países Bajos.

el artículo 8 del Convenio Europeo de Derechos Humanos (CEDH) y con los artículos 7 y 8 de Carta de Derechos Fundamentales de la Unión Europea (CFREU), así como con el Reglamento General de Protección de Datos (RGPD), por infringir el derecho a la privacidad (Van Kordelaar, 2020).

Los sentenciadores, en definitiva, establecen que existe una responsabilidad especial en el uso de las nuevas tecnologías, concluyendo que SyRI supone un incumplimiento del artículo 8 del Convenio Europeo de Derechos Humanos —CEDH—. Es decir, del derecho al respeto a la vida privada y familiar (Van Kordelaar, 2020).

> La polémica, desde un punto de vista jurídico, gira en torno a la falta de transparencia del modelo de riesgos del algoritmo y del uso sesgado de este instrumento, utilizado exclusivamente en barrios donde viven personas con rentas bajas o zonas donde residen personas pertenecientes a minorías. (The Technolawgist, 2020, p.)

Sin perjuicio de la importancia de combatir el fraude en la seguridad social, especialmente porque es pagada por los contribuyentes, para el Tribunal, la falta de transparencia en las nuevas tecnologías conduce a una menor confianza en el Gobierno por parte de los ciudadanos (Van Kordelaar, 2020).

En su razonamiento estima que el artículo 8 del Convenio Europeo de Derechos Humanos —CEDH— obliga a los Países Bajos, como Estado miembro, a equilibrar cuidadosamente los beneficios de las nuevas tecnologías con el derecho a la privacidad. Hecho que no logra hacer mediante la utilización de SyRI al ser incompatible con dicha disposición e innecesario en una sociedad democrática o proporcional (Van Kordelaar, 2020).

Los sentenciadores consideran, así mismo, que deben tener en cuenta la jurisprudencia del Tribunal Europeo de Derechos Humanos —TEDH— a la hora de determinar el alcance del artículo 8 del Convenio Europeo de Derechos Humanos —CEDH—, que ha incluido el respeto por la autonomía y el desarrollo personal bajo el mandato de la citada disposición legal. Agregan que el derecho a la privacidad se encuentra protegido a nivel de la Unión Europea por los artículos 7 y 8 de la Carta de Derechos Fundamentales de la Unión Europea —CFREU—, así como por el Reglamento General de Protección de Datos —RGPD— (Van Kordelaar, 2020).

Mientras que los demandantes reprochan que SyRI utiliza el aprendizaje

profundo y la recopilación de grandes conjuntos de datos para perfilar a las personas, el Estado afirma que esto no es efectivo, ya que el programa simplemente utilizaría árboles de decisión (Van Kordelaar, 2020).

Sin perjuicio de que el artículo 65 de la Ley SUWI permite el aprendizaje profundo y la extracción de datos, de la prueba rendida el Tribunal se logra establecer que el secretismo en torno a SyRI no permite determinar el alcance de sus operaciones (Van Kordelaar, 2020).

Atendido lo anterior, y teniendo en cuenta el principio de transparencia, el principio de limitación de la finalidad y el principio de minimización de datos, el fallo termina por concluir que la legislación SyRI no es suficientemente clara y, por tanto, no es necesaria ni proporcionada en virtud del artículo 8 del Convenio Europeo de Derechos Humanos —CEDH—. Dado que ni el Tribunal ni los particulares pueden conocer su proceso de toma de decisiones, lo que eventualmente daría lugar a discriminación indirecta y a sesgos inesperados que los individuos no tienen cómo cuestionar ni remediar (Van Kordelaar, 2020).

Además, la gran cantidad de datos que pueden ser tratados en virtud del artículo 65 de la Ley SUWI significa que también se infringe el principio de minimización de datos (Van Kordelaar, 2020).

> El uso del SyRI tan solo en determinadas zonas de forma absolutamente discriminatoria por ser percibidas como barrios cuya población tiene una mayor predisposición al fraude fiscal o de la seguridad social, refuerza la estigmatización de estos colectivos limitando cada vez más sus posibilidades. (The Technolawgist, 2020, p.)

"Una de las lecciones más interesantes de esta sentencia es el fundamento jurídico utilizado para establecer la ilegalidad del SyRI". El argumento esperado para sostener aquello habría sido que el uso del algoritmo infringe lo dispuesto en el artículo 22 del Reglamento General de Protección de Datos —RGPD—, el cual establece: "todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar" (The Technolawgist, 2020, p.)

Empero,

> [...] la vulneración del derecho a la vida privada planteaba una vía mucho más clara para establecer la falta de legalidad de la norma.

> Como señala Philip Alston, *Special Rapporteur* de Naciones Unidas sobre cuestiones de extrema pobreza y derechos humanos en el informe presentado durante el proceso, nos encontramos frente al primer juicio del que se tenga conocimiento, que utiliza como fundamento jurídico el incumplimiento de un derecho humano para limitar el uso de determinadas tecnologías, con lo cual se abre una nueva línea argumental para futuros casos. (The Technolawgist, 2020, p.)

3. Conclusiones

A partir del desarrollo de nuestra investigación, hemos podido extraer las siguientes premisas para responder al problema científico planteado en este trabajo y demostrar la efectividad de nuestra hipótesis. La premisa mayor nos dice que el ser humano, por naturaleza, tiende en sus procesos intelectuales al sesgo, mientras que la premisa menor indica que la IA persigue recrear la inteligencia humana. De esto podemos inferir, en consecuencia, que la IA necesariamente va a replicar esa característica. De ahí que el título de nuestro trabajo haga referencia al refrán "de tal palo, tal astilla". No podría ser de otra forma.

Si a eso le sumamos que la IA no solo pretende la semejanza con su creador, sino que es este quien, además, la programa y alimenta de información, la cuestión entra en un círculo vicioso que se ve agravado por el hecho de que las máquinas carecen del contexto y la sensibilidad, que sí tienen los humanos, para eventualmente reconocer y reparar el sesgo.

Según demuestran los resultados del informe de las Naciones Unidas sobre el sesgo de género al que hicimos alusión en nuestra introducción, la situación es preocupante, toda vez que, en promedio, el 88,5 % de la población presenta algún tipo de sesgo hacia las mujeres. Cuestión que, al proyectarse a la IA, que por sus características tiene un efecto amplificador casi ilimitado, puede causar daños catastróficos.

Si bien se han aplicado medidas correctivas, estas no han sido suficientes y, aun cuando somos de la idea de que no puede prescindirse de ninguna de ellas, nos parece que las más adecuadas son aquellas que dicen tener una relación con la creación de algoritmos transparentes alimentados, por medio de datos inocuos con implementación de mecanismos de supervisión y corrección de sesgos, tanto tecnológicos como humanos.

Respecto a estos últimos, es importante que la intervención humana se haga mediante especialistas de

diversas áreas, pero siempre con enfoque de género. No basta con que se incorporen más mujeres al sistema, pues las estadísticas demuestran que sus opiniones son casi tan sesgadas en contra de sus propias congéneres como las de los hombres. Cuestión de la que podemos dar fe, pues en nuestra experiencia hemos sufrido mayor discriminación de parte de las mujeres que de los hombres en materia laboral, financiera y consumeril.

Y, al hablar de perspectiva de género, hay que tener claro que en nuestros días ya no podemos referirnos al clásico dualismo hombre-mujer, sino que debemos incluir a aquellas personas que se consideran no binarias, a fin de evitar la discriminación de las minorías LGTBIQ+.

En último término, es indispensable que se adopten providencias legislativas tanto a nivel nacional como internacional, para la prevención, la auditoria, la corrección y la sanción de los sesgos en la IA. No solo a nivel de género, sino también de otras materias, como son la raza, la religión y la condición social de los individuos, así como la implementación de políticas de capacitación de los usuarios para el conocimiento de estas tecnologías y la autoprotección de sus datos y derechos.

Referencias

Almagro Blanco, P. (2018). Caja negra, caja blanca la inteligencia artificial explicable, inteligencia artificial. El mundo que viene. *Revista de Occidente.* https://www.researchgate.net/publication/328951906_Caja_negra_Caja_blanca_la_inteligencia_artificial_explicable

Alston, P. (2019). Brief by the United Nations Special Rapporteur on extreme poverty and human rights as amicus Curiae in the case of NJCMc.s. De Staat der Nederlanden (SyRI) before the District Court of The Hague (case number; C/09/550982/HAZA18/388). Disponible en: https://www.ohchr.org/Documents/Issues/Poverty/Amicusfinalversionsigned.pdf

Asociación para el Progreso de la Dirección. España. (2021). *Métodos y técnicas de inteligencia artificial: ¿cuáles son y para qué se usan?* Disponible en: https://www.apd.es/tecnicas-de-la-inteligencia-artificial-cuales-son-y-para-que-se-utilizan/#:~:text=15%20t%C3%A9cnicas%20de%20la%20Inteligencia%20Artificial.%20Los%20campos,permitan%20a%20los%20ordenadores%20aprender%20por%20s%C3%AD%20

Benítez Eyzaguirre, L. (2019). Ética y transparencia para la detección de sesgos algorítmicos de género. *Estu-*

dios sobre el Mensaje Periodístico, 25(3), pp.1307-1320. Disponible en: https://dx.doi.org/10.5209/esmp.66989

Cognodata. (2019). Programar inteligencia artificial algoritmos y lenguajes. Disponible en: https://www.cognodata.com/blog/programar-inteligencia-artificial-algoritmos-y-lenguajes

Comisión Europea. (2020a). Libro Blanco. Disponible en: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

Comisión Europea. (2020b). Una Unión de la igualdad: Estrategia para la Igualdad de Género 2020-2025. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A52020DC0152

District Court of The Hague. (2020). NJCMc.s./De Staat der Nederlanden (SyRI), C/09/550982/HAZA18/388). Disponible en: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI%3ANL%3ARBDHA%3A2020%3A865

Faliero, J. (2021, mayo 18). *Inteligencia artificial, sesgo de género algorítmico y violencia de género digital*, Instituto de Estudios Judiciales S.C.B.A. Disponible en: https://www.youtube.com/watch?v=c6gfYaxqLmo

Ferrante, E. (2021). Inteligencia artificial y sesgos algorítmicos ¿Por qué deberían importarnos? *Nueva Sociedad*, N° 294, julio-agosto, pp.27-36. Disponible en: https://www.nuso.org/articulo/inteligencia-artificial-y-sesgos-algoritmicos/

Friedler, S., Scheidegger, C. y Venkatasubramanian, S. (2021) The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Communications of the ACM*, Vol. 64, N° 4, pp.136-143. https://cacm.acm.org/magazines/2021/4/251365-the-impossibility-of-fairness/fulltext

Hao, K. (2019). Auge y caída de las distintas técnicas de IA a lo largo de la historia. *MIT Technology Review*. Disponible en: https://www.technologyreview.es/s/10907/auge-y-caida-de-las-distintas-tecnicas-de-ia-lo-largo-de-la-historia

KeepCoding. (2022). *La inteligencia artificial y los algoritmos.* Disponible en: https://keepcoding.io/blog/inteligencia-artificial-y-los-algoritmos/

Lamas, M. (2000). Diferencias de sexo, género y diferencia sexual. *Revista Nueva Época*, volumen 7, N° 18, enero-abril. Disponible en: https://mediateca.inah.gob.mx/islandora_74/islandora/object/articulo:1058

Lamas, M. (2012). El género es cultura. Disponible en: http://www.paginaspersonales.unam.mx/app/webroot/files/981/El_genero_es_cultura_Martha_Lamas.pdf

Macchiavelli, N. (2021). Perspectiva de género en las nuevas tecnologías. El problema de los sesgos. *Diario Suplemento Derecho y Tecnología*, N° 84. Disponible en: https://dpicuantico.com/sitio/wp-content/uploads/2021/06/Doctrina-Suplemento-Derecho-y-Tecnolog%C3%ADa-07.06.2021-Macchiavelli.pdf

Manes, F. (2020). Sesgos cognitivos: La manera en que pensamos y percibimos el mundo. Disponible en: https://www.diariopopular.com.ar/salud/sesgos-cognitivos-la-manera-que-pensamos-y-percibimos-elmundo-n494497

Manes, F. y Niro, M. (2014). *Usar el cerebro. Conocer nuestra mente para vivir mejor.* Paidós.

National Geographic España. (2019). *Breve historia visual de la inteligencia artificial.* Disponible en: https://www.nationalgeographic.com.es/ciencia/breve-historia-visual-inteligencia-artificial_14419

Organización de las Naciones Unidas —ONU—. (2020). *Human Development Perspectives Tackling Social Norms. A game changer for gender inequalities.* Disponible en: http://hdr.undp.org/sites/default/files/hd_perspectives_gsni.pdf

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura—UNESCO—. (2019a). *Preliminary study on the ethics of artificial intelligence.* Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000367823

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura—UNESCO—. (2019b). *Steering AI and Advanced ICTs for Knowledge Societies.* Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000368711

Prieto, R., Herrera, A., Pérez, J. t Padrón-Godínez, A. (2000, octubre 20), El Modelo Neuronal de Mcculloch y Pitts. Interpretación Comparativa del Modelo. XV Congreso Nacional de Instrumentación, Guadalajara Jalisco-México. Disponible en: https://www.researchgate.net/publication/343141076_EL_MODELO_NEURONAL_DE_McCULLOCH_Y_PITTS_Interpretacion_Comparativa_del_Modelo

Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo. (27 de abril, 2016). Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?qid=1532348683434&uri=-CELEX%3A02016R0679-20160504

Rivas, J. y Espinoza, A. (2020). Sesgo de género e inteligencia artificial. Una perspectiva de Derechos Humanos. *Estado Diario*. Disponible en: https://estadodiario.com/al-aire/sesgo-de-genero-e-inteligencia-artificial-una-perspectiva-de-derechos-humanos/

Sesgo algorítmico. (2022, 21 de enero). En Wikipedia. Disponible en: https://

es.wikipedia.org/wiki/Sesgo_algor%C3%ADtmico

The Technolawgist. (2020). *Derechos humanos en un mundo de algoritmos: la sentencia histórica que ata en corto la implantación de modelos opacos.* Disponible en: https://www.thetechnolawgist.com/2020/02/12/derechos-humanos-en-un-mundo-de-algoritmos-la-sentencia-historica-que-ata-en-corto-la-implantacion-de-modelos-opacos/

Van Kordelaar, S. (2020). Case Netherlands, District Court of the Hague, 5 February 2020 C-09-550982. Fundamental Rights In Courts and Regulation Proyect. Disponible en: https://www.fricore.eu/db/cases/netherlands-district-court-hague-5-february-2020-c-09-550982