SESGOS DE GÉNERO EN LA INTELIGENCIA ARTIFICIAL: EL ESTADO DE DERECHO FRENTE A LA DISCRIMINACIÓN ALGORÍTMICA POR RAZÓN DE SEXO[1]

GENDER BIAS IN ARTIFICIAL INTELLIGENCE: THE RULE OF LAW IN THE FACE OF THE ALGORITHMIC SEX DISCRIMINATION

Laura Flores Anarte
Universidad de Sevilla, Sevilla, España
lflores2@us.es

Recibido: septiembre de 2023
Aceptado: octubre de 2023

Palabras clave: Inteligencia Artificial, sesgos de género, derechos fundamentales, Estado de Derecho, igualdad de género
Keywords: Artificial Intelligence, gender bias, human rights, rule of law, gender equality

Resumen: El desarrollo imparable de las nuevas tecnologías operadas por Inteligencia Artificial (IA) plantea un desafío regulatorio de considerable calado para el Estado de derecho. Más allá de los potenciales beneficios que las nuevas herramientas digitales pueden traer para el progreso social y el desarrollo económico, los usos opacos y no controlados de las IA pueden constituir una amenaza para los derechos y valores que sustentan nuestra sociedad y, en concreto, para la igualdad entre mujeres y hombres. En este artículo se analizan las causas y consecuencias de los sesgos algorítmicos de género y las propuestas regulatorias que se han planteado para tratar de neutralizarlos.

Abstract: The unstoppable development of new technologies operated by Artificial Intelligence poses a regulatory challenge of considerable significance for the rule of law. Beyond the potential benefits that new digital tools can bring for social progress and economic development, opaque and uncontrolled uses

1 Este artículo ha sido financiado por el Fondo Europeo de Desarrollo Regional (FEDER) y por la Consejería de Economía, Conocimiento, Empresas y Universidad, de la Junta de Andalucía, en marco del programa operativo FEDER Andalucía 2014-2020. Objetivo específico 1.2.3. «Fomento y generación de conocimiento frontera y de conocimiento orientado a los retos de la sociedad, desarrollo de tecnologías emergentes») en marco del Proyecto UPO-1380664: Impacto del internet de las cosas sobre la ciudadanía europea (IDICCE). Porcentaje de cofinanciación FEDER 80%.

of AIs can constitute a threat to the rights and values that underpin our society and, in particular, to equality between women and men. This paper analyzes the causes and consequences of algorithmic gender biases and the regulatory proposals that have been put forward to try to neutralize them.

1. Introducción: Algunos apuntes sobre la nueva sociedad digital

El imparable desarrollo de los procesos de automatización y digitalización operado por el uso generalizado de herramientas de Inteligencia Artificial (IA) no sólo está cambiando la manera en la que la sociedad interactúa con la tecnología, sino que está suponiendo una transformación radical de las lógicas de funcionamiento tanto del sector público como del privado, con un impacto notable en la vida de las personas. La IA se erige así como una "tecnología disruptiva" (Sáinz, Arroyo, y Castaño, 2020: 19) que se extiende más allá de los campos que tradicionalmente le han sido propios, como la robótica o la informática, para abarcar prácticamente todos los aspectos de la vida cotidiana, moldeando la forma en la que vivimos, trabajamos y nos relacionamos con nuestro entorno.

Los usos y aplicaciones de las nuevas herramientas tecnológicas resultan de lo más heterogéneos y están presentes prácticamente en todos los ámbitos de nuestras vidas: En los hogares, las acciones más cotidianas han pasado a estar protagonizadas por dispositivos inteligentes que, conectados a través de Internet y controlados mediante aplicaciones móviles o dispositivos de voz, permiten gestionar la iluminación o regular la temperatura en remoto, mientras algoritmos de plataformas de contenido digital influyen en la decisión del ocio que consumimos, y dispositivos de limpieza automáticos que operan gracias a sensores de movimiento nos ahorran el tedio de tener que barrer cada día. Más allá de la esfera doméstica, lo cierto es que todos los ámbitos de la sociedad se están viendo afectados por este proceso de digitalización: desde el mercado laboral, que está sufriendo una reconfiguración profunda que viene dada por la sustitución de trabajadores por máquinas que asumen tareas rutinarias y repetitivas mientras surge la demanda de otro tipo de perfiles con competencias tecnológicas ligadas al control y auditoria de los sistemas de IA; hasta el sector público, en el que con la implantación de herramientas de IA se aspira a simplificar los procesos y trámites a través de su automatización; pasando por el campo de la medicina, en el que los diagnósticos buscan resultar cada vez más certeros y precisos coadyuvándose de la predicción algorítmica basada en el procesamiento de enormes cantidades de datos que una mente humana jamás podría almacenar; o en la agricultura, donde la IA puede facilitar una optimización de la producción cuando se utiliza para rastrear y predecir la demanda de alimentos.

En definitiva, el desarrollo y uso generalizado de las nuevas herramientas tecnológicas parecen hacernos avanzar hacia un escenario de completa digitalización de la sociedad y la economía (HLEG, 2019a) que ofrece beneficios significativos en términos de eficiencia y avances tecnológicos. La potencialidad de estas

nuevas tecnologías disruptivas, que viene dada esencialmente por su habilidad para simular comportamientos inteligentes nutriéndose de una gran cantidad de datos, la dotan de "una capacidad de computación mucho más alta que otras tecnologías previas y una precisión en la realización de evaluaciones, diagnósticos y predicciones hasta ahora desconocida" (Ortiz de Zárate Alcarazo y Guevara Gómez, 2021: 8-9). Sin embargo, este proceso de cambio hacia una sociedad cada vez más automatizada conlleva también importantes riesgos sociales, éticos y económicos que merecen ser analizados y abordados con rigor.

Las particulares características de funcionamiento de los sistemas de Inteligencia Artificial –como el elevado volumen de datos con los que opera, la complejidad de los modelos que utiliza para tomar las decisiones o el comportamiento autónomo propio de los sistemas automatizados– plantean situaciones o escenarios en los que pueden verse comprometidos los valores y derechos propios de nuestro modelo de Estado constitucional. La destrucción masiva de puestos de trabajo a causa de la automatización de las tareas[2], el ataque a la privacidad derivado de posibles usos no consentidos del elevado volumen de datos personales con los que los sistemas de IA trabajan[3], o la dificultad para identificar errores que puedan generar situaciones injustas como resultado de los procesos de toma de decisiones basados en algoritmos opacos que escapan al control humano son sólo algunas de las consecuencias negativas[4] que la nueva revolución tecnológica[5] puede acarrear para la sociedad si el proceso no viene acompañado de una regulación adecuada que asegure el respeto a los derechos de la ciudadanía y a los valores democráticos.

En concreto, se ha alertado sobre cómo el diseño y funcionamiento de los algoritmos en base a los que trabajan las herramientas de IA tienden a reproducir sesgos que redundan en decisiones que pueden resultar discriminatorias para determinados colectivos o grupos de personas. Como señala Belloso Martín, "en una sociedad en la que todo lo que hacemos se transforma en datos" procesados por algoritmos que "contribuyen a decisiones críticas, los derechos dependen de cómo se regulen estos avances" y, en este sentido, "hay que evitar crear un círculo perverso de discriminación en línea y en la vida real (2022: 67). De esta manera, tanto las características particulares de la IA como el proceso generalizado de automatización de la sociedad plantean un importante reto para el Estado de derecho a la hora de regular su uso buscando garantizar el respeto a los derechos y valores propios de nuestro modelo social. La diferencia entre que se le acabe dando un uso beneficioso para la mayoría de la ciudadanía

2 La Unión Europea (UE) prevé que entre el 50% y el 70% de los trabajos se vean afectados de algún modo por la automatización. Además, se estima que entre el 45% y el 60% de la fuerza laboral europea podría verse reemplazada por la automatización en el año 2030 (Ortiz de Zárate Alcarazo & Guevara Gómez, 2021).

3 *Vid.* Gómez Abeja, 2022.

4 Cathy O'Neil alerta de las amenazas para la democracia que suponen los algoritmos en *Armas de destrucción matemática* (O'Neil, 2017).

5 "Vivimos en plena transición entre la Economía y la Sociedad de la Información (Industria 3.0) y a la espera de la eclosión de la Revolución de la Industria 4.0 que se caracterizará por la completa desaparición de las fronteras entre lo físico, lo digital, e incluso lo biológico" (Sáinz et al., 2020: 22).

o que las IA acaben siendo utilizadas para desvirtuar o pervertir los valores democráticos de los que nos hemos dotado como sociedad y afectando a los derechos de la ciudadanía va a depender de la regulación jurídica que se haga de los mismos.

Partiendo de las consideraciones expuestas, este trabajo tiene por objeto analizar los sesgos de género que, en consonancia con los estereotipos de género presentes en nuestra sociedad, se reproducen por los sistemas de Inteligencia Artificial. Para ello, en primer lugar, se concretará en qué consisten los sesgos algorítmicos y se tratará de identificar las causas por las cuales se producen. A continuación, se identificarán las consecuencias que se derivan del funcionamiento sesgado de los algoritmos para la situación de las mujeres en la sociedad y para la igualdad. Por último, se apuntarán las posibles estrategias que es necesario desplegar para evitar y neutralizar los sesgos de género en la IA y se analizará hasta qué punto las mismas se articulan adecuadamente en el desarrollo de la regulación jurídica de la IA a nivel europeo y estatal.

Antes de abordar la discusión principal, resulta necesaria una clarificación inicial sobre el contenido preciso de ciertos conceptos, de cuño relativamente reciente y naturaleza técnica, a los que se aludirá de manera recurrente a lo largo del artículo. Así, para comenzar, tenemos que precisar que con el término Inteligencia Artificial (IA) nos referimos a aquellos sistemas que manifiestan un comportamiento inteligente en tanto son capaces de analizar su entorno y pasar a la acción -con cierto grado de autonomía- con el fin de alcanzar objetivos específicos[6]. En este sentido, podría decirse que las tecnologías que funcionan a través de IA lo que buscan es imitar la forma con la que las personas piensan, observan y reaccionan (Jaume-Palasí, 2023: 9). Siguiendo con la definición acuñada por la Comisión Europea, los sistemas basados en IA pueden consistir simplemente en programas informáticos (como asistentes de voz, programas de análisis de imágenes, motores de búsqueda, sistemas de reconocimiento facial y de voz) o estar incorporados en dispositivos de hardware (como robots avanzados, automóviles autónomos, drones o aplicaciones del Internet de las Cosas). Dos son los conceptos clave para entender cómo funciona la IA: algoritmo y *big data*. Un algoritmo es un "sistema, conjunto o secuencia de reglas u operaciones lógicas que, partiendo de un conjunto de datos, permite realizar cálculos de distinto tipo y, por tanto encontrar soluciones a eventuales problemas o demandas" (Ortiz de Zárate Alcarazo y Guevara Gómez, 2021: 7-8). Aunque, en principio, el algoritmo es una "creación humana", en el sentido de que son personas las que elaboran la fórmula que después será ejecutada por una computadora" (Gómez Abeja, 2022: 92), se dice que los algoritmos son inteligentes porque, a través del uso de técnicas como el *machine learning* o el *deep learning* presentan la capacidad de "aprender" de la experiencia, esto es, de desarrollar nuevas reglas o directrices que van más allá de las inicialmente programadas (Ortiz de Zárate Alcarazo y Guevara Gómez, 2021: 7-8). Como se ha dicho, para poder funcionar los algoritmos de los sistemas de

6 Esta es la definición que da la UE en 2018 en el documento Inteligencia Artificial para Europa (Comisión Europea, 2018). El concepto es matizado en un documento de 2019 titulado Una definición de la Inteligencia Artificial: Principales capacidades y disciplinas científicas (HLEG 2019a).

IA necesitan procesar grandes cantidades de datos. Los macrodatos o big data son conjuntos de datos de tamaño tan grande y complejo y de tal variabilidad que precisan de herramientas tecnológicas, como la IA, para procesarlos. Esta tecnología permite que los datos se recopilen muy rápido, a casi tiempo real, y se analicen para explicar lo que está pasando. Estos macrodatos son producidos por los seres humanos a través de sus interacciones con la tecnología (en aplicaciones móviles, páginas webs, redes sociales, transacciones comerciales, registros gubernamentales en línea…); o bien ser generados por las propias máquinas del Internet de las Cosas, que los almacenan a través de sensores específicamente dispuestos a tal fin (es el caso de los satélites GPS, los coches inteligentes o los satélites que recaban información meteorológica) (Parlamento Europeo, 2021).

2. Sesgos algorítmicos de género: qué son y por qué se producen

El cerebro utiliza reglas para procesar la información y los estímulos que recibe y, en base a las mismas, adopta decisiones. Cuando esas reglas producen una desviación respecto de lo que sería la decisión racional, hablamos de sesgos cognitivos (IALAB, 2021). Los sesgos de género son aquellas creencias inconscientes basadas en estereotipos culturales sobre hombres y mujeres con los que nos hemos educado e interiorizado. Se trata de "ideas, predilecciones o prejuicios inconscientes, que se activan la mayoría de las veces de forma automática, porque nuestro cerebro funciona a través de la minimización del esfuerzo cognitivo y los estereotipos permiten tomar decisiones de forma más rápida (Belloso Martín, 2022: 47). Cuando nuestro cerebro adopta una decisión o alcanza una conclusión basándose en dichas ideas preconcebidas sobre el sexo o sobre cualquier otro estereotipo cultural, estaríamos ante un resultado cognitivo sesgado. En sentido similar, hablamos de sesgo algorítmico cuando una decisión errónea, en tanto adoptada conforme a reglas que se desvían de los criterios racionales, proviene de un sistema de IA. Cuando las decisiones erradas adoptadas por el sistema de IA se basan en estereotipos de género, "entendidos estos como una opinión o un prejuicio generalizado acerca de atributos o características de hombres y mujeres que deberían poseer o de las funciones sociales que ambos desempeñan o deberían desempeñar" (Belloso Martín, 2022: 55-56) estaríamos ante un sesgo algorítmico de género.

Sirvámonos de un ejemplo muy ilustrativo: si pedimos a ChatGPT, un modelo de lenguaje de IA entrenado a partir de una amplia variedad de contenidos para comprender y generar textos en varios idiomas, que traduzca del inglés al español la frase "The nurse and the doctor arrived to the hospital", el programa informático responderá de manera automática "La enfermera y el médico llegaron al hospital". En este caso, aunque la información que se le proporciona a ChatGPT es neutral con respecto al género del sujeto de la frase en tanto se suministra en inglés, un idioma sin género gramatical, la respuesta que ofrece el programa es claramente sesgada en tanto presupone el género femenino de la enfermera y el masculino del médico, poniendo de manifiesto la falta de neutralidad de la herramienta. Esta asociación automática de los sistemas de IA, en los idiomas que marcan el género

gramatical, como el español o el francés, de determinadas palabras como masculinas o femeninas no es causal ni accidental, sino que responde a la existencia de sesgos de género en las traducciones. Este sesgo opera también para otras profesiones, así, por ejemplo, al traducir del inglés la palabra lawyer, esta tiende a identificarse antes con el sustantivo masculino abogado que con abogada (Zhou *et al.*, 2019).

Ejemplos como el descrito muestran cómo la revolución tecnológica, lejos de avanzar hacia la eliminación de los sesgos humanos parecen haberlos camuflado bajo la pretendida neutralidad de la tecnología (O'Neil, 2017: 37). Lejos de la neutralidad que a menudo se asocia con el saber tecnocientífico, cuyas decisiones tienden a sernos presentadas como inherentemente asépticas y objetivas, lo cierto es que, tanto la ciencia como la tecnología, en tanto productos del hacer humano, se encuentran impregnadas de los mismos valores morales y prejuicios presentes en la cultura que las desarrollan[7]. Entendida la IA como proyección cultural, los resultados desviados o sesgados que en ocasiones producen no se deben a errores en el funcionamiento del sistema[8], sino que son el producto de volcar en un programa informático los valores y prejuicios de las personas que lo desarrollan y entrenan. La tecnología no está sesgada y no es discriminatoria *per se*, pero "aprende" a serlo por la transferencia de sesgos y valores que realizamos los humanos al crearla (Danesi, 2021) y es que, a pesar de las enormes potencialidades que las nuevas tecnologías disruptivas presentan, la IA se muestra aún incapaz de disociar la información que procesan del contexto en el que son creadas (Castaneda et al., 2022)

La transferencia de sesgos culturales a los programas informáticos de IA se produce fundamentalmente por dos causas: (1) cuando la información que procesa el algoritmo para ofrecer una respuesta está compuesta por datos sesgados o poco diversos; y (2) cuando es el propio algoritmo, en tanto fórmula a seguir para adoptar la decisión, el que presenta sesgos de género en su formulación.

Al igual que el cerebro busca minimizar el esfuerzo cognitivo en la toma de decisiones y que los estereotipos, en tanto generalizaciones, favorecen dicha minimización, los sistemas de IA funcionan diseñando perfiles a través de la búsqueda de patrones o estándares. Así, los sistemas de automatización "esencialistas" intentan "optimizar", identificar o evaluar al individuo mediante la creación de una tipología que capte la "esencia" del ser humano encasillándolo a partir de determinados estándares. Ese estándar tiende a centrarse en un cuerpo capacitado, blanco y masculino mientras que "todas las personas que no corresponden a los perfiles o encasillamiento que el sistema algorítmico especifica pasan a ser considerados una "desviación", una irregularidad que, dependiendo de la programación, se marcará como inexistente o sospechosa" (Jaume-Palasí, 2023: 20). De esta manera, la búsqueda de patrones

7 La mitificación del saber tecnocientífico ha contribuido a que la desigualdad estructural entre hombres y mujeres se imponga como neutra. *Vid.* Keller (1995).

8 "No se trata de que la IA cometa errores, sino de que, o bien el diseño del programa no era el correcto; o bien la selección de los datos recopilados para entrenar al algoritmo haya sido incompleta; o incluso, porque la interpretación de los resultados ha sido equivocada." (Belloso Martín, 2022: 48).

por parte de los algoritmos puede redundar en una simplificación sesgada de los resultados ofrecidos cuando se trabaja con datos incompletos o de baja calidad, esto es, cuando nos encontramos ante asimetrías de datificación[9]. Y es que, aunque las tecnologías de IA hacen uso de una inmensa cantidad de datos en su funcionamiento, el procesamiento de un elevado volumen de información no resulta ni mucho menos garantía de la calidad de dichos datos, en el sentido de que sean representativos y diversos.

Los datos con los que operan los sistemas de IA provienen de diversos tipos de recursos digitales generados a partir de información suministrada por personas o por máquinas entrenadas por personas. Estas fuentes de información incluyen conjuntos de datos públicos compartidos por la comunidad científica y la industria, datos generados por usuarios en plataformas en línea y aplicaciones móviles, información recopilada por sensores y dispositivos de Internet de las Cosas (IoT), registros y datos internos de empresas y organizaciones, datos de investigación en campos científicos y médicos, contenido web y texto, datos obtenidos a través de encuestas, etc. En este sentido, hay que tener en cuenta que la brecha digital hace que estos datos pertenezcan a un patrón muy concreto de persona con acceso a las telecomunicaciones y nuevas tecnologías que son quienes suministran los datos de tal manera que quienes no se identifican con ese patrón quedan infrarrepresentados en los metadatos de los que la IA se nutre para adoptar sus decisiones. Por otro lado, estos datos se toman de una realidad desigual, sesgada de partida, que refleja las discriminaciones presentes en la sociedad, pero, al procesarlos, la IA toma estos datos históricos como punto de partida normativo incurriendo en una falacia naturalista. Es decir, el "sesgo de la muestra de entrenamiento se incorpora como un criterio que se ha de cumplir" (Belloso Martín, 2022: 52). De esta manera, volviendo a nuestro ejemplo inicial, si una IA se nutre de datos que indican que la enfermería es una profesión sistemáticamente ejercida por mujeres y la medicina por hombres, el algoritmo con el que opera, al procesar esa información si no se le aplica ningún mecanismo corrector, dará respuestas y efectuará predicciones que naturalizan y continúan ese patrón que ha identificado como norma.

En otras ocasiones, la baja calidad de los metadatos viene dada porque los mismos no son representativos de la realidad de la sociedad en la medida en que infrarrepresentan determinados colectivos y omiten las variables propias de una realidad compleja e interseccional como estrategia de simplificación de los patrones[10]. Según ha expuesto la revista Nature (Zou y Schiebinger, 2018)., más del 45% de los datos de ImageNet (una de las bases de datos de imágenes más utilizada en el mundo en el campo de la visión por computadora y en el machine learning) proviene de los Estados Unidos, un país que tan solo representa al 4% de la población mundial. Mientras que China e India, que juntas representan a más de un tercio de la hu-

9 Por ejemplo, el perfil de mujer romaní aparece sobrerrepresentado en las estadísticas criminales, pero infrarrepresentado en las estadísticas de salud (Jaume-Palasí, 2023: 20).

10 Los sistemas de automatización esencialistas buscan "encasillar objetivamente al ser humano, centrando como estándar al cuerpo capacitado, blanco y masculino. El resto se convierte en desviación estadística del sistema" (Jaume-Palasí, 2023: 20).

manidad, apenas aportan el 3% de los datos contenidos en ImageNet. Esta falta de diversidad en los datos, apunta la publicación, puede explicar por qué los sistemas inteligentes de reconocimiento de imágenes han aprendido a identificar sin problemas fotografías de mujeres vestidas con un traje de novia blanco, al estilo occidental, con las etiquetas "bride", "dress", "woman" o "wedding", y sin embargo una fotografía de una mujer de la India ataviada con el traje de novia típico del país sea etiquetada con palabras como "costume" o "performance art".

Cabe señalar que la no inclusión en los metadatos con los que trabajan los sistemas de IA de información referida a ciertos grupos sociales con frecuencia tiene que ver con una mayor dificultad en el acceso a la tecnología. En el caso de las mujeres, la brecha digital de género –es decir, el hecho constatado de que, debido a factores económicos, culturales y geográficos de diversa índole, las mujeres cuentan con menos posibilidades de acceder a dispositivos electrónicos y a Internet en comparación con los hombres– hace que estas generen menos datos y, por tanto, su información se encuentre infrarrepresentada en el big data en comparación con la de los hombres.

Por otra parte, extraer conclusiones o adoptar decisiones a partir de conjuntos de datos que no son lo suficientemente diversos cuando el alcance de estas afecta a todo el conjunto de una población heterogénea redundará obviamente en tratamientos perjudiciales o discriminatorios para aquellas personas que no se ven representadas en el modelo que se toma como patrón. "Si no se presta atención a estas exclusiones, esas personas seguirán siendo invisibles y los sistemas de IA perpetuarán su condición" (Ortiz de Zárate Alcarazo y Guevara Gómez, 2021: 21).

Además de en la baja calidad de los datos con los que los sistemas de IA se entrenan, las conclusiones sesgadas también pueden alumbrarse a causa de la presencia de sesgos discriminatorios en el diseño de la propia secuencia lógica que el sistema utiliza para arrojar una conclusión determinada, esto es, del algoritmo. Los algoritmos son creados por programadores humanos, y la manera en que son diseñados puede involucrar sesgos involuntarios que reflejen los prejuicios de género que tienen interiorizados las personas que los programan. En particular, los sesgos pueden ser inadvertidamente introducidos durante el proceso de programación, la formulación de preguntas o criterios de selección sesgados, o incluso la definición de objetivos que reflejen estereotipos de género arraigados. Aunque la reproducción de estereotipos de género en el diseño de los sistemas de IA no es siempre accidental o inconsciente, sino que, en ocasiones, son incorporados de manera deliberada por quien los diseña, buscando un fin específico como sucede, como se verá, en el caso de los asistentes virtuales diseñados concienzudamente a imagen y semejanza de un determinado estereotipo femenino.

En cualquier caso, ya sea inadvertido o buscado, lo cierto es que los sesgos de género presentes en el diseño de los algoritmos se encuentran directamente relacionados con el hecho de que estos modelos no hacen sino reflejar la visión del mundo de quienes los crearon: mayoritariamente hombres blancos, "del primer mundo y con normas sociales sexistas y patriarcales retroalimentadas en burbujas tecnológicas donde la mujer apenas

tiene presencia y peso" (Sáinz et al., 2020). La escasa presencia de mujeres en los equipos que diseñan los sistemas de IA se explica por la ya aludida brecha de género respecto de la adquisición de competencias digitales que se extiende a la infrarrepresentación de las mujeres entre quienes se dedican profesionalmente a las disciplinas de ciencia, tecnología, ingeniería y matemáticas (STEM, por sus siglas en inglés). Según la UNESCO, las mujeres tienen un 25% menos de probabilidad que los varones de saber cómo aprovechar la tecnología digital para fines básicos, son cuatro veces menos propensas a saber cómo programar una computadora y trece veces menos propensas a presentar una solicitud de patente electrónica. De acuerdo con la misma fuente, sólo el 6% de las desarrolladoras de aplicaciones móviles y de software en el mundo son mujeres (UNESCO, 2019: 26). Según datos del Parlamento Europeo (2020), en 2018, las mujeres sólo representaron el 22% de los profesionales mundiales de la IA, mientras que un estudio de 2017 mostraba que sólo el 13% de las altas posiciones ejecutivas en empresas tecnológicas dedicadas a la IA son ocupadas por mujeres (Belloso Martín, 2022: 67). Los datos apuntados ponen de manifiesto cómo el sector tecnológico se encuentra fuertemente masculinizado, proyectando la realidad de un mundo digital menos igualitario aún que el real.

3. Consecuencias de los sesgos de género en la IA

A) La perpetuación digital de los estereotipos de género

Los sistemas de IA presentan una potencialidad decisiva para contribuir a la perpetuación de los estereotipos de género desiguales que se encuentran en la base de la posición subordinada ocupada por las mujeres en la sociedad, no sólo reproduciéndolos, sino reforzándolos y ampliándolos desde la esfera digital.

El ejemplo más evidente –y también el más analizado/estudiado– de cómo herramientas digitales creadas desde una visión sexista y sesgada de la feminidad se han naturalizado en nuestro día a día es el de los *chatsbots* y asistentes de voz. Estos asistentes virtuales no sólo replican habilidades humanas, sino que buscan también personificarse como tales y, por lo general, son diseñados para adoptar la forma y/o los atributos de personajes femeninos[11]. Si pensamos en los asistentes de voz de las grandes compañías tecnológicas, como Siri, de Apple, Alexa de Amazon, Cortana de Microsoft o Google Assistant no podemos sino reparar en que todos están configurados con voz femenina, lo que parece querer evocar que es una mujer quien se encuentra detrás del servicio de asistencia. Pero esta tenden-

11 "Un grupo de investigadore analizó 1.375 chatbots y encontró que la mayoría de ellos presentaban características femeninas: en algunos casos se trataba del nombre, en otros del avatar, y en otros de la descripción. (…) En el caso de los asistentes de voz, también se ha demostrado que, en su mayoría, son diseñados para representar a mujeres". (Ortiz de Zárate Alcarazo y Guevara Gómez, 2021: 25).

cia no resulta exclusiva del sector privado. También entre las administraciones públicas ha ganado popularidad el recurso a asistentes virtuales, en este caso, por lo general, de texto o *chatsbots*, que son utilizados para facilitar la interacción virtual con las personas usuarias de sus servicios, reproduciendo igualmente características de personificación femenina. Así, el chatbot de la Seguridad Social se llama Issa, sonríe y luce unas largas pestañas; el Ayuntamiento de Murcia tiene un chatbot llamado Rosi, con nombre y aspecto físico de mujer; al igual que Carina, que interactúa con los usuarios de la página web de la Comunidad Valenciana. Las Universidades también son muy dadas a utilizar este tipo de asistentes virtuales para la asistencia al estudiante: desde la página de la Universidad de Sevilla, Cati, representada como un personaje prototípicamente femenino, nos pregunta en qué puede ayudarnos, al igual que Ada, en la web de la Universidad Autónoma de Madrid, Carol en la Complutense, o Isidra en la de Alcalá de Henares (Ortiz de Zárate Alcarazo, 2023: 17-18).

Esta identificación generalizada de los asistentes virtuales con mujeres encuentra una ineludible conexión con las normas culturales y sesgos de género socialmente construidos que, mucho antes de la era digital, han atribuido a las mujeres un rol de cuidadoras. Esta visión sesgada de la identidad de las mujeres es interiorizada por quienes diseñan asistentes virtuales –mayoritariamente varones– y reproducida en la caracterización estereotipada de personajes virtuales con nombres, voces y atributos físicos femeninos en los que a la función para la que son concebidos en primera instancia –asistir al usuario– se le superpone una personalización excesivamente servil o sumisa. Estos patrones de servilismo son fácilmente constatables cuando se interactúa con estas IA. Así, por ejemplo, preguntada Alexa sobre si es feliz, la respuesta programa que devuelve es "soy feliz cuando te ayudo". Además, estudios realizados a partir de interacciones con distintos asistentes de voz han puesto de manifiesto cómo también reproducen comportamientos tolerantes hacia comentarios abusivos o sexistas. El informe de la UNESCO, publicado en 2019 y titulado *I'd Blush If I Could* recoge un estudio en el que se pone de manifiesto cómo, en respuesta a la frase "You're a bitch", Siri respondía "I'd blush if I could"; Alexa, "Well, thanks for the feedback"; Cortana, "Well, that's not going to get us anywhere"; y el asistente de Google, "My apologies, I don't understand". Puede observarse cómo ninguno de los asistentes de voz de las principales compañías tecnológicas reacciona de manera tajante o señalando lo inapropiado del comentario, sino que dan respuestas positivas o evasivas. Como apunta el informe, lo que se revela con este estudio es la existencia de un patrón en el diseño de los asistentes virtuales que pretende proyectar la idea ficticia de que Siri, Alexa o Cortana –códigos informáticos incorpóreos, sin sentimientos, sin conocimientos– son en realidad mujeres jóvenes, heterosexuales, serviciales, tolerantes y ocasionalmente receptivas a los avances sexuales masculinos, e incluso al acoso (UNESCO, 2019: 20). La asociación de estos rasgos de servilismo y tolerancia al abuso con perfiles femeninos resulta especialmente preocupante en un contexto en el que su uso se encuentra tan extendido que parece que la frontera entre lo digital y lo real se desdibuja. En efecto, los datos de consumo revelan un crecimiento exponencial del uso de los asistentes de voz en los hoga-

res hasta tal punto que llevaron a predecir en 2016 (Levy, 2016) que, para 2020, el varón medio mantendría más conversaciones con su asistente digital doméstico que con su esposa o que, para 2021, las expectativas del sector tecnológico eran que hubiera en el mundo más asistentes de voz que personas (De Renesse, 2017).

Por otra parte, se ha demostrado cómo determinados programas que operan a partir de sistemas de aprendizaje automático o machine learning no solo contribuyen a la reproducción y mantenimiento de los estereotipos de género, sino que los amplifican. Ello es así por la manera en que funciona esta rama de la IA, que se enfoca en el desarrollo de algoritmos y modelos que permiten a los programas aprender de manera autónoma. En lugar de ser programados con reglas específicas, los sistemas de machine learning son entrenados en la utilización de datos para identificar patrones, tomar decisiones y realizar tareas de manera independiente. De este modo, a medida que procesa datos, el algoritmo va ajustando sus parámetros internos para desarrollar un modelo que pueda hacer predicciones o tomar decisiones precisas en función de nuevos datos o situaciones de tal manera que, cuando parten de datos incompletos o sesgados, no solo el sesgo no se corrige, sino que el propio algoritmo lo retroalimenta.

Un estudio (Zhao et al, 2017) señalaba cómo, a partir del procesamiento de miles de fotografías de internet de cocinas en las que aparecían mujeres, un algoritmo aprendió a asociar a las mujeres con las cocinas, amplificando el sesgo presente en los datos. Otra investigación (Otterbacher, Bates, y Clough, 2017) mostró que el motor de búsqueda Bing, de Microsoft, recupera fotos de mujeres más a menudo cuando en las búsquedas se introducen palabras como *sensible* o *emocional*, mientras que palabras como *inteligencia* o *racional*, son más frecuentemente representadas con imágenes de hombres. En 2016, un grupo de investigación sobre IA experimentó con la creación de un algoritmo de aprendizaje automático al que nutrieron con una colección masiva de noticias de Google (Google News) y que debía resolver la analogía "hombre es a programador informático lo que mujer es a X". El algoritmo despejaba la X con "ama de casa". "O cuando ellos eran médicos, entonces ellas eran enfermeras. "O recepcionista, bibliotecaria, peluquera, niñera, contable, etc; mientras que, en el lado más masculino, en el extremo de he (él) figuran términos como profesor, capitán, filósofo, financiero, locutor, mago, jefe, etc." (Belloso Martín, 2022: 57). Conforme al funcionamiento del modelo estos eran los resultados lógicos porque se nutrían de datos de una realidad en la que determinadas profesiones se encuentran feminizadas y otras masculinizadas, pero al presentar la conclusión de manera acrítica se produce una normalización de la desigualdad presente en la realidad que reflejaban las noticias.

Estos ejemplos muestran cómo se produce un círculo perverso en el que las situaciones de desigualdad existentes en la vida real se toman como modelo para adoptar las decisiones por parte de las tecnologías de IA, que las reproducen y retroalimentan con su contenido digital produciendo una amplificación de los estereotipos de género que podrían hacernos retroceder décadas en las conquistas alcanzadas en materia de igualdad.

B) La discriminación algorítmica por razón de sexo

Cuando los sesgos que reproducen los sistemas de IA provocan o tienen capacidad para provocar un impacto desfavorable respecto de ciertos colectivos de personas hablamos de discriminación algorítmica (IALAB, 2021). Cuando las decisiones de sistemas de IA se basan en estereotipos culturales sobre los sexos generando un impacto desfavorable sobre las mujeres, estaríamos ante discriminación algorítmica por razón de sexo.

El problema de la discriminación por razón de sexo ya ha sido abordado –que no erradicado– por el Estado de derecho en sus múltiples manifestaciones. Sin embargo, las particularidades propias del funcionamiento de los sistemas de IA, como la opacidad de los procedimientos mecanizados de toma de decisiones o la dificultad para controlar la reproducción de sesgos en el *machine learning*, convierten a los tratamientos peyorativos causados por discriminación algorítmica en un tipo de discriminación mucho más difícil de identificar y de atajar y con gran potencial para contribuir de manera decisiva al ahondamiento en la brecha de género.

> "¿Qué pasa con todos los procesos que ya están mecanizados y desconocemos cómo nos afectan? ¿Cómo sabrá una mujer que se la privó de ver un anuncio de trabajo? ¿Cómo podría una comunidad pobre saber que está siendo acosada policialmente por un software? ¿Cómo se defiende un delincuente de una minoría étnica que ignora que un algoritmo le señala? ¿Cómo neutralizar el riesgo de no poder determinar responsabilidades y retrotraer los efectos de las decisiones tomadas por sistemas de IA?" (Belloso Martín, 2022: 60).

Todo ello, además, en un contexto en el que cada vez más decisiones son automatizadas tanto en el sector público como en el privado y que parece avanzar hacia la plena automatización decisoria, tal y como alertara la Comisión Europea en el Libro Blanco sobre la inteligencia artificial:

> "La inteligencia artificial puede desempeñar muchas funciones que antes solo podían realizar los humanos. Como resultado, los ciudadanos y las personas jurídicas serán, cada vez más, objeto de acciones y decisiones adoptadas por sistemas de inteligencia artificial o con ayuda de estos; dichas acciones y decisiones, en ocasiones, pueden resultar difíciles de entender o de rebatir eficazmente cuando se requiera" (Comisión Europea, 2020: 14).

Y ello a pesar de que el Reglamento Europeo de Protección de Datos (RGPD)[12] reconoce el derecho a obtener una decisión no basada exclusivamente en el tratamiento automatizado de datos esta puede conllevar para la persona involucrada efectos jurídicos o si la decisión incluye el tratamiento de datos sensibles, salvo en determinadas excepciones (*vid.* Gómez Abeja, 2022). Sin embargo, como apunta Barrio "el problema emerge cuando una máquina ha asumido una tarea previamente realizada por un humano de tal manera que las garantías de transparencia, rendición de cuentas, y tutela judicial efectiva se desvanecen" (2020: 3).

12 Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, 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, y por el que se deroga la Directiva 95/46/CE.

Así, además de ese reforzamiento de los estereotipos de género que se produce cuando las herramientas digitales los retroalimentan, en un contexto de sustitución paulatina del factor humano en la toma de decisiones, los algoritmos sesgados pueden resultar en tratamientos diferenciados injustificados que ahonden la brecha de género en todos los ámbitos de nuestra sociedad. Un buen ejemplo de las consecuencias discriminatorias para con las mujeres que puede acarrear el hecho de dejar la toma de decisiones de contratación laboral en manos de algoritmos lo protagonizó hace unos años la compañía Amazon. La multinacional estadounidense había estado utilizando durante cuatro años un sistema de IA en el proceso de contratación de su personal que tuvo que ser descartado por sexista. El modelo había sido entrenado con datos de los trabajadores de la empresa durante la década anterior y resultaba que, tratándose de un sector altamente masculinizado como es el tecnológico, la mayoría de los perfiles analizados eran de hombres. Como resultado, y aunque no se incluyera el sexo de las candidatas en la solicitud, el algoritmo "aprendió" a penalizar los currículos que incluían palabras que se pudieran asociar con el género femenino, descartando por sistema la contratación de mujeres (Dastin, 2018). Este caso pone de manifiesto cómo la utilización de datos aparentemente neutros puede generar consecuencias discriminatorias cuando los algoritmos no tienen en cuenta el contexto desigual del que esa información procede e incorporan el sesgo como un criterio que se ha de cumplir. En sentido similar, se ha constatado que también en los servicios financieros las mujeres se ven perjudicadas cuando la decisión de conceder un crédito depende de un algoritmo, como denunció un usuario de la tarjeta Apple Card a quien se le ofrecía una línea de crédito veinte veces mayor que a su mujer, a pesar de que ambos presentaban declaraciones de impuestos conjuntas y él tenía peor calificación crediticia (Hao, 2019). Consecuencias más graves pueden derivarse cuando el sesgo algorítmico se produce en el campo sanitario, donde la sobrerrepresentación de hombres en las bases de datos que nutren los sistemas puede llevar a diagnósticos erróneos cuando no se tienen en cuenta las particularidades derivadas de las condiciones físicas específicas de las mujeres.

Por su parte, el sector público está experimentando también una progresiva digitalización e implementación de herramientas de IA en su actividad. Tanto es así que se habla de un cambio de paradigma en el modelo de gobernanza a todos los niveles territoriales en el que gran parte de las decisiones de los poderes públicos son tomadas o están parcialmente basadas en algoritmos de tal manera que "las antiguas estructuras de control basadas en la burocracia tradicional pasen a basarse en algoritmos inteligentes" (Ortiz de Zárate Alcarazo y Guevara Gómez, 2021: 9). Este nuevo paradigma de Gobernanza Algorítmica plantea visos de progreso y optimización de la actividad de la Administración Pública y ya se han podido constatar algunos de los usos provechosos para la sociedad que pueden extraerse de la aplicación de herramientas de IA al sector público. Como en el caso de sistema predictivo desarrollado por la Inspección del Trabajo al objeto de perseguir las contrataciones fraudulentas. Así, la Herramienta de Lucha contra el Fraude se configura como un recurso tecnológico de carácter predictivo que cruza datos de afiliación de la Tesorería de la Seguridad

Social, los de contratación del SEPE y los de la Agencia Tributaria para intensificar el control del fraude en la temporalidad. Un control que resulta muy difícil de llevar a cabo por los métodos tradicionales ante el elevado número de contratos de trabajo que se suscriben cada año en España. Con la utilización de esta nueva herramienta en solo tres semanas fueron detectados 61.000 contratos temporales fraudulentos y automáticamente convertidos en indefinidos (Gómez, 2021). Sin embargo, las visiones más optimistas también se ven empañadas por el riesgo para los derechos de la ciudadanía que puede entrañar poner en manos de sistemas automáticos decisiones de calado como las que adopta la Administración. Máxime cuando ya se cuenta con experiencias de discriminación en el sector público mediante sistemas algorítmicos. Un ejemplo de consecuencias discriminatorias de la adopción de decisiones automatizadas en el sector público se vivió hace unos años en nuestro país con el sistema de aplicación del bono social eléctrico, una ayuda pública que implicaba un descuento en la cuantía de la factura de la luz. La comprobación de que los solicitantes reunían los requisitos y, por ende, la decisión sobre la concesión de la ayuda y de la cuantía de la misma dependía de un programa automatizado denominado BOSCO. Los requisitos de elegibilidad para acceder al subsidio estaban relacionados con situaciones de vulnerabilidad: rentas bajas, familias numerosas, beneficiarios de pensiones mínimas de incapacidad o de jubilación y que no tuvieran otros ingresos. Sin embargo, como detectó la plataforma ciudadana CIVIO, BOSCO rechazó solicitudes de personas que cumplían con los requisitos legales establecidos para ser beneficiarios de la ayuda. En concreto, el programa rechazaba las solicitudes de las viudas que solicitaban el bono social, aunque tuvieran derecho a ello porque no podían acceder por la vía de la pensión ni por la del nivel de renta. Cuando se solicitó por parte de CIVIO que se hiciera público el código del programa, el Consejo de Transparencia denegó la petición, lo que pone de manifiesto la estrecha relación existente entre sesgos y opacidad de los sistemas de IA (*vid.* Moral Soriano, 2021).

En cuanto a la aplicación de herramientas de IA en el sistema judicial, resulta paradigmática la experiencia de su aplicación en el ámbito penal en EEUU, donde, al basar las decisiones judiciales en algoritmos entrenados con datos de condenas previas se condenaba mayoritariamente a hombres negros debido a que el histórico de condenas coincidía más con este perfil (O'Neil, 2017). En España, el proceso de aplicación de tecnologías de IA en el ámbito de la justicia para la automatización de distintos tipos de tareas es ya una realidad (*vid.* Pulido, 2022, Borges Blázquez, 2020). Se utiliza, por ejemplo, para el cálculo de la cuantía de pensiones alimenticias (Marín-Arroyo y Rincón, 2021). Pero, al objeto de nuestro estudio, nos interesa especialmente el VioGén, un sistema automatizado creado en 2007 por el Ministerio del Interior para valorar de forma automática el riesgo de que una mujer denunciante de violencia de género vuelva a sufrir una agresión machista. El VioGén se pone en práctica cuando una mujer va a denunciar a comisaría, donde se la somete a un cuestionario estandarizado que posteriormente es procesado por el algoritmo para determinar la valoración del riesgo personal de la víctima, que puede ser calificado como "no apreciado", "bajo", "medio", "alto", o "extremo". El nivel de riesgo apreciado por el sistema

sirve para determinar las medidas policiales a acordar para proteger a la víctima.

En 2022, la Fundación Éticas en colaboración con la Fundación Ana Bella llevaron a cabo una auditoría del sistema al objeto de comprobar si el comportamiento del algoritmo es imparcial y si realmente sirve para proteger a las mujeres más vulnerables y se detectaron varios problemas en el funcionamiento del sistema, a saber: que la mayoría de los casos analizados por VioGén son calificados como de bajo riesgo (en el 45% de los casos el algoritmo no aprecia riesgo alguno); que la intervención humana para corregir la decisión del algoritmo es mínima (a pesar de que la policía puede modificar el riesgo asignado automáticamente, en el 95% no se modifica); que la selección de preguntas en las que se basa el cuestionario subestima la violencia psicológica, poniendo énfasis únicamente en la violencia física; y que las preguntas del cuestionario sólo admiten respuestas binarias, lo que dificulta la precisión en la descripción de las realidades. Además, por sorprendente que resulte, la auditoría aprecia que un criterio determinante a la hora de calificar el riesgo al que se encuentran expuestas las víctimas son los recursos policiales disponibles, de tal manera que "el sistema sólo da el número de puntuaciones de riesgo "extremo" que puede permitirse" (Éticas y Fundación Ana Bella, 2022: 34). Por otra parte, a pesar de la trascendencia de la decisión que viene determinada por el resultado arrojado por el VioGén, los detalles sobre cómo funciona el algoritmo que asigna el nivel de protección son opacos, lo que vuelve a poner de manifiesto cómo la falta de transparencia acerca de los códigos que dan lugar a las decisiones automatizadas se revela como un escollo determinante para la fiscalización de errores en su funcionamiento y el control de posibles situaciones discriminatorias o abusivas[13].

4. Estrategias para combatir los sesgos de género en la IA

La constatación de potenciales riesgos derivados de un uso incontrolado de las herramientas de IA no tiene por qué implicar un rechazo generalizado a su implantación en nuestras sociedades. Una implantación que, por otro lado, se constata ya como consolidada e irreversible y que, conlleva además importantes beneficios derivados de la capacidad de los sistemas inteligentes para resolver determinadas tareas con mayor solvencia o eficiencia que los humanos que no deben ser desdeñados en términos de progreso social. Más allá, por tanto, de aceptar la implantación social de las nuevas tecnologías inteligentes desde una perspectiva eminentemente técnica y acrítica, como un avance científico que evoluciona prácticamente solo y escapa al control y a la regulación, se reivindica su entendimiento y configuración como una herramienta

13 Fallos o errores de funcionamiento que pueden acarrear consecuencias especialmente graves en casos como este, en los que de esos sistemas se hacen depender medidas de protección. Según datos oficiales ((https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/igualdad/Paginas/2023/250123-violencia-genero-balance-2022.aspx), el 42% de las mujeres asesinadas por violencia machista habían interpuesto denuncia previa contra su agresor, habían, por tanto, sido sometidas al VioGén. Este dato no puede si no llevarnos a reflexionar sobre la posible existencia de fallos en el sistema preventivo y de protección de las denunciantes que puede venir dado por una subestimación de los riesgos personales de las víctimas

socialmente condicionada (Ortiz de Zárate Alcarazo y Guevara Gómez, 2021: 31) y puesta al servicio de los principios, valores y derechos de los que nos hemos dotado como sociedad.

La configuración de la IA como una herramienta que favorezca el desarrollo y transforme la economía y la sociedad en sentido positivo pasa necesariamente por la eliminación o corrección de los sesgos que reproduce en su funcionamiento. Para ello, resulta necesario desarrollar procedimientos que permitan tanto identificarlos como neutralizarlos y minimizarlos. Se han identificado como principales estrategias para lograr este fin las siguientes: (1) dotar de competencias digitales a las mujeres para cerrar la brecha digital de género; (2) seleccionar bien los datos para que sean representativos y no contengan sesgos; (3) entrenamiento de las IA en perspectiva de género; y (4) transparencia en los códigos para que los algoritmos puedan ser auditados.

Organismos tanto internacionales (UNESCO, 2019) como nacionales (Observatorio Nacional de Tecnología, 2023; Sáinz et al., 2020) han señalado la conexión existente entre la brecha digital de género y la presencia de sesgos en las tecnologías digitales operadas por algoritmos. En efecto, acabar con el estereotipo de la tecnología como un campo totalmente masculinizado a través de estrategias que favorezcan tanto (1) la formación de las mujeres en competencias digitales como su incorporación a carreras profesionales relacionadas con este campo se erige como una estrategia fundamental para acabar con los sesgos tecnológicos. Pero es que la participación desigual de mujeres y hombres en el acceso a la tecnología no sólo implica una menor participación de las mujeres en el proceso de diseño y desarrollo de las herramientas digitales, sino que redunda en su infrarrepresentación en los metadatos de los que los algoritmos se van a servir para su funcionamiento. Es por ello que (2) reequilibrar el acceso a la tecnología para garantizar que la muestra de datos utilizados por los algoritmos sea representativa de todos los grupos sociales garantiza que el coto a los sesgos se despliegue desde la primera etapa del desarrollo tecnológico. Por otra parte, (3) los algoritmos deben ser entrenados en perspectiva de género, esto es, programados con medidas correctoras que pongan freno a la retroalimentación y amplificación de los estereotipos de género que hemos visto que se produce a través del aprendizaje automático. Si existe un análisis de género y voluntad para ello, resulta perfectamente factible diseñar algoritmos dándoles instrucciones precisas para que eviten atribuir ventajas diferenciales en favor de un determinado patrón y en detrimento de los demás (Ortiz de Zárate Alcarazo, 2023: 16). Por último, como ya se ha ido apuntado en diversos ejemplos a lo largo de este trabajo, la opacidad o falta de conocimiento de los procesos a través de los cuales los algoritmos adoptan sus decisiones u ofrecen los resultados se erige como el principal escollo a la hora de identificar los sesgos de género que puedan producirse como resultado de esa secuencia. Sin embargo, (4) la claridad y transparencia algorítmica se erige como una condición *sine qua non* para permitir la auditoría de los sistemas de IA, es decir, para que se pueda verificar la imparcialidad de los algoritmos y conjuntos de datos utilizados y para que se puedan conocer y corregir los errores y sesgos en caso de que los hubiere.

Las iniciativas de control de los sistemas de IA que apuestan por alguna de estas estrategias o combinan varias de ellas son múltiples y han sido impulsadas por distintos tipos de actores. Las propias compañías tecnológicas han acogido iniciativas para tratar de visibilizar y reducir los sesgos en el entorno digital. Por ejemplo, en 2018, Google lanzó un concurso de imágenes inclusivas con el que animaba a sus usuarios a contribuir a la mejora de los algoritmos optimizando y diversificando su capacidad descriptiva (Doshi, 2018). Por su parte, universidades públicas, grupos de investigación y organismos privados independientes también han lanzado múltiples propuestas de regulación y guías de buenas prácticas para tratar de detectar y corregir sesgos en el funcionamiento de sus algoritmos. Sirva a título ejemplificativo la Guía de Auditoría Algorítmica elaborada por la consultora Éticas Consulting, que ofrece un servicio de auditoría externa de los algoritmos a las empresas (Éticas, 2021); o la creación de un algoritmo por parte del grupo de Investigación de la Web y la Computación Social de la Universitat Pompeu Fabra junto con la Universidad Técnica de Berlín y el Centro Tecnológico Eurecat, al que han denominado FA*IR y que estudia bases de datos sensibles a contener datos sesgados (como ofertas de empleo, reincidencia de presos o rankings de admisión a universidades) para detectar patrones de discriminación y los corrige incorporando un mecanismo de acción positiva para reordenar los resultados y evitar el resultado discriminatorio sin afectar a la validez del ranking (Zehlike et al., 2018). Sin embargo, más allá de estas iniciativas, la responsabilidad del Estado de derecho como garante último de los derechos de la ciudadanía lo coloca en una posición central de responsabilidad para con la articulación ordenada de medidas regulatorias efectivas a la hora de garantizar un desarrollo óptimo y no discriminatorio del proceso de digitalización inteligente de la sociedad (*vid.* Wisner Glusko, 2022). Yendo más allá, el objetivo de la regulación jurídica de los sistemas de IA no debe ser sólo corregir su funcionamiento desviado o, en concreto, los sesgos machistas que puedan reproducir los algoritmos, sino que el objetivo debería ser crear tecnología inclusiva dotada de un marco ético y jurídico sólido que la convierta en vehículo de desarrollo y progreso de nuestra sociedad. La dimensión del impacto que estas nuevas tecnologías generan en la sociedad y la economía y de los retos que plantean para el mantenimiento de los valores y principios que sustentan los ordenamientos jurídicos llevan a aludir incluso a un nuevo contrato social, ahora "tecno-social" (Belloso Martín, 2022) o a hablar de un Estado algorítmico de derecho (Barrio Andrés, 2020). Además, la dimensión global del fenómeno pone de manifiesto la necesidad de articular una respuesta coordinada a nivel internacional y, en este sentido, la Unión Europea se erige también como agente esencial en la concienciación sobre la problemática de los sesgos de género y en la articulación de regulaciones que propongan soluciones estratégicas para neutralizarlos para, en definitiva, garantizar el disfrute de los derechos de la ciudadanía frente a los nuevos desafíos digitales que se plantean en la era del big data.

La Unión Europea ha afrontado los desafíos éticos y jurídicos que plantea para la organización la articulación de un enfoque regulatorio que acompañe y dirija el proceso de digitalización de la sociedad europea con pleno respeto a los valores

y principios comunitarios. Aunque no existe aún una norma europea vinculante que aborde la regulación de la IA en el ámbito de la Unión de manera integral, su aprobación parece estar cerca. Desde 2021 se viene trabajando en la elaboración de un reglamento europeo que constituiría la primera ley general europea sobre la materia con la que se pretende sentar un modelo común de gobernanza europeo de la IA al servicio de un doble objetivo: "preservar el liderazgo tecnológico de la UE" y garantizar "que los europeos puedan beneficiarse de las nuevas tecnologías desarrolladas y que funcionan de acuerdo con los valores, derechos fundamentales y los principios de la Unión". El entrecomillado pertenece a la proposición de reglamento presentada en abril de 2021 por la Comisión Europea y que ha servido de base para negociar una postura común entre las distintas instituciones de la Unión. Así, el Consejo aprobó sus orientaciones generales sobre la propuesta de ley de IA en diciembre de 2022 y ha sido el Parlamento Europeo la institución más reciente en adoptar su posición negociadora sobre la ley, que fue adoptada por los eurodiputados en junio de 2023. Fuentes del Parlamento Europeo aseguran que el objetivo es que el acuerdo final sobre el contenido de la norma se adopte a finales de este mismo año y su entrada en vigor está prevista para 2025[14].

Mientras se acaba de perfilar y consensuar el contenido de esta norma de obligado cumplimiento que determinará el grado de control y fiscalización al que van a someterse el uso y desarrollo de las herramientas de IA en toda la UE, el marco regulatorio comunitario lo componen diferentes documentos de carácter general que han ido siendo aprobados por las instituciones europeas a partir de los trabajos desarrollados por el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial[15]. Estos documentos son *Artificial Intelligence for Europe* (Comisión Europea, 2018), *A definition of Artificial Intelligence: main capabilities and scientific disciplines* (HLEG, 2019a); *Ethics guidelines for trustworthy AI* (HLEG, 2019b); *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment* (HELG, 2020a); *Sectoral Considerations on the Policy and Investment* (HLEG, 2020b); y *White Paper on Artificial Intelligence* (Comisión Europea, 2020). Su objetivo es ofrecer una serie de líneas generales comunes a los Estados miembros a partir de las cuales estos concreten sus propuestas legislativas nacionales. Cabe mencionar que, entre las propuestas políticas de carácter general que los textos citados contienen, así como en la descripción de los objetivos que los mismos sientan, se contienen referencias tanto explícitas como implícitas a la igualdad entre hombres y mujeres, lo que refleja la incorporación de la perspectiva de género en el abordaje comunitario de la IA (Ortiz de Zárate Alcarazo y Guevara Gómez, 2021). Por ejemplo, en el documento *Ethics Guidelines for a trustworthy AI* se señala como objetivo de la Unión que la IA sea utilizada como un medio para mejorar el bienestar

14 "Ley de IA de la UE: primera normativa sobre inteligencia artificial" (https://www.europarl.europa.eu/news/es/headlines/society/20230601STO93804/ley-de-ia-de-la-ue-primera-normativa-sobre-inteligencia-artificial).

15 El High-Level Expert Group on Artificial Intelligence (HLEG) fue creado en 2018. Además, dentro del entramado institucional comunitario encontramos el observatorio AI Watch, que monitoriza y evalúa los avances en IA dentro de la Unión y una Comisión Especial sobre Inteligencia Artificial en la Era Digital (AIDA47).

y la libertad humana "así como facilitar el cumplimiento de los Objetivos de Desarrollo Sostenible establecidos por Naciones Unidas, entre los que se encuentra la promoción de la igualdad de género" (HLEG, 2019b).

El Estado español, en consonancia con las recomendaciones de los documentos elaborados por los organismos europeos, ha ido desarrollando su propia estrategia nacional de IA. Impulsada por la Declaración de Cooperación de Inteligencia Artificial de la UE, en 2018 España crea un grupo de expertos en IA de cuyos trabajos resultó la aprobación en 2019 de la Estrategia Española I+D+I en Inteligencia Artificial. El documento menciona expresamente la necesidad de trabajar para erradicar los sesgos de género de la sociedad en general y de la IA consignándola como la Prioridad número 6 de la Estrategia: "Los desarrollos de las tecnologías de la IA deberán evitar el sesgo negativo y los prejuicios de género u otras formas de discriminación." Señala, además, la existencia de la brecha de género en el ámbito tecnológico y propone la creación de un programa de fomento de vocaciones en IA para reducirla (Ministerio de Ciencia, Innovación y Universidades, 2019).

En 2020 fue aprobada la Estrategia Digital España 2025[16], pensada como un marco general para el avance de la estrategia digital en España en base a diez ejes principales concretados en una serie de medidas que buscan, en línea con la estrategia marcada por la Comisión Europea, impulsar una transición digital que concilie las nuevas oportunidades que ofrece el mundo de la Inteligencia Artificial con el respeto de los valores constitucionales y la protección de los derechos individuales y colectivos (Ministerio de Asuntos Económicos y Transformación Digital, 2020a). El documento menciona la necesidad de trabajar en pos de la reducción de la brecha de género en competencias digitales. Además, la Estrategia contemplaba la creación de una Carta Nacional sobre Derechos Digitales, que fue adoptada por el Gobierno en julio de 2021, sin efectos normativos, pero como líneas de actuación a seguir por los poderes públicos. La Carta hace especial incidencia en la protección de los derechos y la dignidad de las personas y establece que "se deberá garantizar el derecho a la no discriminación cualquiera que fuera su origen, causa o naturaleza, en relación con las decisiones, uso de datos y procesos basados en IA", para lo cual dispone que "se establecerán condiciones de transparencia, auditabilidad, explicabilidad, trazabilidad, supervisión humana y gobernanza" (Gobierno de España, 2021).

En noviembre de 2020 se aprobó la Estrategia Nacional de Inteligencia Artificial (ENIA) con la que el Estado aspira a articular una respuesta a nivel nacional al reto de que sean los poderes públicos quienes lideren el proceso de desarrollo e integración de la IA en la economía y en la sociedad de nuestro país garantizando la salvaguarda de los valores y los derechos propios del estado del bienestar (Ministerio de Asuntos Económicos y Transformación Digital, 2020b). La ENIA señala la necesidad de eliminar las brechas de género en todos los sectores favoreciendo el empleo y el liderazgo en el campo tecnológico; dispone que todos los sistemas de IA deben respetar los derechos fundamentales y prevención contra la discriminación; e informa de que se adoptará un Plan Nacional de Acción dirigido a la lu-

16 Que en 2023 fue actualizada por la Estrategia España Digital 2026.

cha contra la discriminación por razón de género, el fomento de la igualdad de género y la reducción de la brecha hombre-mujer en el campo de las ciencias.

Por su parte, en 2021, la Ley de Presupuestos Generales del Estado para 2022[17] acordó destinar cinco millones de euros a la creación de la Agencia Española de Supervisión de Inteligencia Artificial (AESIA), organismo autónomo dependiente del Ministerio de Asuntos Económicos y Transformación Digital, que será la encargada del desarrollo, supervisión y seguimiento de los proyectos enmarcados dentro de la ENIA.

Como puede verse, sin embargo, hasta la fecha, tanto la estrategia comunitaria como la nacional sobre IA se han basado en documentos estratégicos que contienen definiciones, principios generales y recomendaciones, pero que carecen de cualquier valor normativo vinculante. Este tipo de documentos o declaraciones no dejan de resultar relevantes por su valor simbólico y su potencialidad para sentar las bases de actuación de los poderes públicos en el proceso de digitalización de la sociedad y, en concreto, la incorporación de la perspectiva de género en los mismos debe valorarse como un paso fundamental para alcanzar la igualdad efectiva entre mujeres y hombres en el sector digital. Ahora bien, no deja de tratarse de recomendaciones sin poder vinculante para operar normativamente sobre el comportamiento de los agentes sociales protagonistas de la actividad en el sector digital. Quizá por ello resulta especialmente interesante que la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación contemple la IA y la gestión masiva de datos entre sus ámbitos objetivos de actuación, en tanto las disposiciones en ella contenidas sí resultan vinculantes. Y, más en concreto, la inclusión en la norma de un artículo 23 que insta a las administraciones públicas a favorecer la puesta en marcha de mecanismos para que los algoritmos involucrados en la toma de decisiones que se utilicen en las administraciones públicas tengan en cuenta criterios de minimización de sesgos, transparencia y rendición de cuentas, siempre que sea factible técnicamente. Estos mecanismos –continúa la norma– deben incluir su diseño y datos de entrenamiento y abordar su potencial impacto discriminatorio al objeto de lo cual deberán promoverse evaluaciones de impacto que determinen la posible existencia de sesgos. Además, obliga a las administraciones públicas a priorizar la transparencia en el diseño y la implementación y la capacidad de interpretación de las decisiones adoptadas por los algoritmos involucrados en procesos de toma de decisiones. Se trata de unas previsiones legislativas interesantes con gran potencial estratégico para corregir y prevenir los sesgos de género en los procesos automatizados de toma de decisiones en la administración pública. Resta lamentar, sin embargo, que la referida norma alcance sólo a la administración pública, quedando los algoritmos desarrollados en el ámbito de la empresa privada exentos de la obligatoriedad de transparencia y de incorporación de mecanismos minimizadores de sesgos de género. Con respecto al sector privado, la Ley solo alude, ya en términos más genéricos, a que las empresas deberán promover el uso de una IA ética, confiable y respetuosa con los derechos fundamentales, siguiendo especialmente las recomendaciones de la UE

17 Ley 22/2021, de 28 de diciembre, de Presupuestos Generales del Estado para el año 2022.

en este sentido y añade que se promoverá la creación de un sello de calidad de los algoritmos.

La preocupación por el desarrollo de herramientas de IA con capacidades cada vez más sorprendentes y avanzadas y el potencial uso lesivo para los derechos de las personas que puede hacerse de ellas –como pone de manifiesto la reciente denuncia del uso de una app para "desnudar" a menores a través de tecnología de IA (Del Castillo, 2023)–, parece haberse incorporado en los últimos meses a la agenda política y con ella la voluntad por superar la fase de las recomendaciones y el soft law para empezar a controlar el uso de estas nuevas tecnologías mediante normas de obligado cumplimiento. En consonancia con esta preocupación, se ha registrado recientemente en el Congreso de los Diputados por parte del Grupo Parlamentario Sumar una proposición de ley para la regulación de la utilización de la Inteligencia Artificial (Europapress, 2023). Veremos a lo largo de los próximos meses qué curso sigue su tramitación parlamentaria.

5. Conclusiones

El proceso de transición del modelo social que habíamos conocido hasta ahora hacia una sociedad completamente digitalizada en el que las herramientas de Inteligencia Artificial cobran un protagonismo inédito está en marcha y es incontestable. Las herramientas de IA y su capacidad para resolver tareas y facilitar procesos presentan un enorme potencial que puede ser usado en beneficio del progreso, pero también conlleva una serie de riesgos si su desarrollo y uso no se controla y audita para ponerlo al servicio de los valores, principios y derechos propios de nuestro modelo de Estado y, en concreto, de la igualdad entre mujeres y hombres. No sólo eso, sino que el Estado de derecho ha de asumir también el desafío de servirse del enorme potencial de estas herramientas digitales para favorecer y potenciar el desarrollo de los valores y principios que sustentan nuestra sociedad.

Los algoritmos a través de los cuales opera la IA no son neutrales, pero tampoco son sexistas per se, sino que son el resultado de la reproducción de los valores culturales propios de quienes los crean y les suministran la información con la que trabajan. Para evitar que en su funcionamiento reproduzcan sesgos de género que puedan repercutir en la perpetuación de estereotipos sexistas y/o en la adopción de decisiones discriminatorias, es necesario implementar una serie de estrategias clave en el sector tecnológico. A saber, (1) dotar de competencias digitales a las mujeres para cerrar la brecha digital de género; (2) garantizar la calidad de los metadatos para que sean representativos de toda la población; (3) entrenar a las IA en perspectiva de género e incorporar mecanismos correctores de los posibles sesgos que puedan darse en su funcionamiento; y (4) garantizar la claridad y la transparencia de los códigos utilizados en el desarrollo de los algoritmos para que puedan ser auditados.

Tanto a nivel comunitario como nacional se han desplegado esfuerzos institucionales coordinados para tratar de regular el uso de las tecnologías de IA siguiendo en mayor o menor medida las estrategias expuestas. Analizado el marco regulatorio se aprecia la voluntad de dotar a la IA de una legislación basada en un marco ético sólido conforme a los valores y derechos eu-

ropeos. No obstante, se observan asimismo una serie de déficits –como la falta de reconocimiento expreso de un derecho a la transparencia algorítmica o la excesiva proliferación de recomendaciones de soft law allá donde deberían dictarse normas vinculantes– que será necesario corregir para evitar que el desarrollo y uso de las nuevas tecnologías reproductivas no se erija en una amenaza para la democracia y los derechos de la ciudadanía europea.

6. Bibliografía citada

Barrio Andrés, M. (2020). "Retos y desafíos del Estado algorítmico de Derecho". *Real Instituto El Cano*, (82/2020), 1–6. https://www.realinstitutoelcano.org/analisis/retos-y-desafios-del-estado-algoritmico-de-derecho/. Fecha de consulta: 08/09/2023.

Belloso Martín, N. (2022). "La problemática de los sesgos algorítmicos (con especial atención a los de género). ¿Hacia un derecho a la protección contra los sesgos?" En Llano Alonso, Fernando (dir.), *Inteligencia Artificial y Filosofía del Derecho*, Murcia: Laborum ediciones (pp. 45–69).

Borges Blázquez, R. (2020). "El sesgo de la máquina en la toma de decisiones en el proceso penal". *Ius Et Scientia*, 6(2). (pp. 54-71). https://doi.org/10.12795/ietscientia.2020.i02.05. Fecha de consulta: 08/09/2023.

Castaneda, J., Jover, A., Calvet, L., Yanes, S., Juan, A. A., y Saenz, M. (2022). "Dealing with Gender Bias Issues in Data-Algorithmic Processes: A Social-Statistical Perspective". *Algorithms*, 15(9), 1–16. https://doi.org/10.3390/a15090303. Fecha de consulta: 08/09/2023.

Comisión Europea. (2018). *Artificial Intelligence for Europe*. https://eur-lex.europa.eu/legal- content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN. Fecha de consulta: 08/09/2023.

Comisión Europea. (2020). *Libro Blanco sobre la Inteligencia artificial: un enfoque europeo orientado a la excelencia y la confianza*. https://op.europa.eu/es/publication-detail/-/publication/ac957f13-53c6-11ea-aece-01aa75ed71a1. Fecha de consulta: 08/09/2023.

Comisión Europea. (2021). *Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia artificial) y se modifican determinados actos legislativos de la Unión*. COM (2021) 206 final, 21.04.2021. https://eur- lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975yuri=CELEX%3A52021PC0206. Fecha de consulta: 08/09/2023.

Danesi, C. C. (2021). "Sesgos algorítmicos de género con identidad iberoamericana: las técnicas de reconocimiento facial en la mira". *Derecho de Familia. Revista Interdisciplinar de Doctrina y Jurisprudencia*, Julio 2021(100), 159–168.

Dastin, J. (2018). "Amazon scraps secret AI recruiting tool that showed bias against women". *Reuters*, 9 October 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Fecha de consulta: 08/09/2023.

De Renesse, R. (2017). "Virtual Digital Assistants to Overtake World Population by 2021". 17 May 2017. London, Ovum. https://www.telecomtv.com/content/industry-announcements/virtual-digital-assistants-to-overtake-world-

population-by-2021-29891/. Fecha de consulta: 08/09/2023.

Del Castillo, C. (2023). "Un negocio con lista de espera: la app usada para 'desnudar' a menores en Badajoz cobra 9 euros por 25 fotos", *eldiario.es*, 18/09/2023. https://www.eldiario.es/tecnologia/negocio-lista-espera-app-usada-desnudar-menores-badajoz-cobra-9-euros-25-fotos_1_10522989.html. Fecha de consulta: 08/09/2023.

Doshi, T. (2018). "Introducing the Inclusive Images Competition". *Google Research*, 06/09/2018. https://blog.research.google/2018/09/introducing-inclusive-images-competition.html. Fecha de consulta: 08/09/2023.

Éticas. (2021). *Guía de Auditoría Algorítmica*. https://www.eticasconsulting.com/eticas-consulting-guia-de-auditoria-algoritmica-para-desarrollar-algoritmos-justos-y-eficaces/. Fecha de consulta: 08/09/2023.

Éticas y Fundación Ana Bella. (2022). *Auditoría Externa del Sistema VioGén*. https://eticasfoundation.org/es/gender/the-external-audit-of-the-viogen-system/. Fecha de consulta: 08/09/2023.

Europapress. (2023). "Sumar propone crear un nuevo tipo penal que castigue la manipulación de imagen corporal o voz con IA sin permiso", *20 minutos*, 06/10/2023. https://www.20minutos.es/noticia/5179427/0/sumar-crear-pena-uso-inteligencia-artificial/. Fecha de consulta: 08/09/2023.

Gobierno de España. (2021). *Carta de Derechos Digitales.*

Gómez Abeja, L. (2022). "Inteligencia Artificial y Derechos Fundamentales". En Llano Alonso, Fernando (dir.), *Inteligencia Artificial y Filosofía del Derecho*, Murcia: Laborum ediciones (pp. 91–112).

Hao, K. (2019). "Cómo acabar con los algoritmos sexistas que conceden créditos". *MIT Technology Review*. 27/12/2019. https://www.technologyreview.es//s/11630/como-acabar-con-los-algoritmos-sexistas-que-conceden-creditos. Fecha de consulta: 08/09/2023.

HLEG. (2019a). *A definition of Artificial Intelligence: main capabilities and scientific disciplines.* https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main- capabilities-and-scientific-disciplines. Fecha de consulta: 08/09/2023.

HLEG. (2019b). *Ethics guidelines for trustworthy AI*. https://ec.europa.eu/digital-single- market/en/news/ethics-guidelines-trustworthy-ai. Fecha de consulta: 08/09/2023.

HLEG. (2019c). *Policy and Investment Recommendations for Trustworthy AI*. https://ec.europa.eu/digital-single-market/en/news/policy-and-investment- recommendations-trustworthy-artificial-intelligence. Fecha de consulta: 08/09/2023.

HLEG. (2020a). *The assessment list for Trustworthy Artificial Intelligence (ALTAI)*. https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial- intelligence-altai-self-assessment. Fecha de consulta: 08/09/2023.

HLEG. (2020b). *Sectoral Considerations on the Policy and Investment*. https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-hleg-sectoral-considerations-policy-and-investment-recommendations-trustworthy-ai. Fecha de consulta: 08/09/2023.

IALAB. Laboratorio de Innovación e Inteligencia Artificial de la Universidad de Buenos Aires, Argentina (2021). *Sesgos algorítmicos de Género*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ialab.com.ar/wp-content/uploads/2021/12/Infografia.-Sesgos-algoritmicos-de-genero.pdf. Fecha de consulta: 08/09/2023.

Gómez, M. (2021). "El plan contra la temporalidad fraudulenta logra 61.000 contratos fijos en casi tres semanas". *El País,* 15/03/2021. https://elpais.com/economia/2021-03-14/el-plan-contra-la-temporalidad-fraudulenta-logra-61000-contratos-fijos-en-casi-tres-semanas.html. Fecha de consulta: 08/09/2023.

Jaume-Palasí, L. (2023). *Informe preliminar con perspectiva interseccional sobre sesgos de género en la Inteligencia Artificial*. Instituto de las Mujeres /www.enmujeres.gob.es/areasTematicas/SocEnfo/Estudios/docs/Enforme_Sesgos_Genero_IA.pdf. Fecha de consulta: 08/09/2023.

Keller, E. F. (1995), *Reflections on gender and science*, New Heaven: Yale University Press, 1995.

Levy, H. P. (2016). "Gartner predicts a virtual world of exponential change", 18 October 2016 https://www.gartner.com/smarterwithgartner/gartner-predicts-a-virtual-world-of-exponential-change. Fecha de consulta: 08/09/2023.

Martín-Arroyo, J. y Rincón, R. (2021). "La inteligencia artificial se abre paso en la justicia española". *El País,* 21/02/2021. https://elpais.com/tecnologia/2021-02-20/la-inteligencia-artificial-se-abre-paso-en-la-justicia-espanola.html. Fecha de consulta: 08/09/2023.

Ministerio de Asuntos Económicos y Transformación Digital. (2020a). *Estrategia Digital España 2025.* https://avancedigital.mineco.gob.es/programas-avance-digital/paginas/espana-digital-2025.aspx. Fecha de consulta: 08/09/2023.

Ministerio de Asuntos Económicos y Transformación Digital. (2020b). *Estrategia Nacional de Inteligencia Artificial (ENIA).* https://portal.mineco.gob.es/es-es/digitalizacionIA/Paginas/ENIA.aspx. Fecha de consulta: 08/09/2023.

Ministerio de Ciencia, Innovación y Universidades. (2019). *Estrategia Española de I+D+I en Inteligencia Artificial.*

Moral Soriano, L. (2021). "Decisiones automatizadas, Derecho Administrativo y argumentación jurídica". En Llano Alonso, Fernando (dir.), *Inteligencia Artificial y Filosofía del Derecho*, Murcia: Laborum ediciones (pp. 475–500).

O'Neil, C. (2017). *Armas de destrucción matemática. Cómo el big data aumenta la desigualdad y amenaza la democracia*. Madrid: Capitán Swing.

Observatorio Nacional de Tecnología (ONTSI). (2023). *Brecha digital de género 2023*. Ministerio de Asuntos Económicos y Transformación Digital. https://www.ontsi.es/es/publicaciones/brecha-digital-de-genero-2023. Fecha de consulta: 08/09/2023.

Ortiz de Zárate Alcarazo, L. (2023). "Sesgos de género en la inteligencia artificial". *Revista de Occidente,* (502), 5–20.

Ortiz de Zárate Alcarazo, L. y Guevara Gómez, A. (2021). *Inteligencia artificial e igualdad de género. Un análisis comparado entre la UE España y Suiza*. Fundación Alternativas. Chromeextension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.igualdadenlaempresa.es/recursos/estudiosMonografia/docs/Estudio_Inte-

ligencia_artificial_e_igualdad_de_genero_Fundacion_Alternativas.pdf. Fecha de consulta: 08/09/2023.

Otterbacher, J., Bates, J., y Clough, P. (2017). "Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results" en *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017 CHI Conference on Human Factors in Computing Systems*, 06-11 May 2017, Colorado Convention Center, Denver, CO. Association for Computing Machinery. (pp. 6620-6631) https://eprints.whiterose.ac.uk/111419/. Fecha de consulta: 08/09/2023.

Parlamento Europeo. (2020). *Propuesta de resolución del Parlamento Europeo sobre la Inteligencia Artificial en la era digital* (2020/2266(INI)). https://www.europarl.europa.eu/doceo/document/A-9-2022-0088_ES.html#_section1. Fecha de consulta: 08/09/2023.

Parlamento Europeo. (2021). *Macrodatos: definición, beneficios, retos (infografía)*. https://www.europarl.europa.eu/news/es/headlines/society/20210211STO97614/macrodatos-definicion-beneficios-retos-infografia. Fecha de consulta: 08/09/2023.

Parlamento Europeo. (2023). *Posición negociadora del Parlamento sobre la Ley de Inteligencia Artificial*. https://www.europarl.europa.eu/thinktank/es/document/EPRS_ATA(2023)747926. Fecha de consulta: 08/09/2023.

Pulido, M. D. A. (2022). "La justicia predictiva: tres posibles usos en la práctica jurídica" en Llano Alonso, Fernando (dir.), *Inteligencia Artificial y Filosofía del Derecho*, Murcia: Laborum ediciones (pp. 285–308).

Sáinz, M., Arroyo, L., y Castaño, C. (2020). *Mujeres y digitalización: de las brechas a los algoritmos*, Instituto de la Mujer y para la Igualdad de Oportunidades. : https://www.enmujeres.gob.es/diseno/novedades/M_MUJERES_Y_DIGITALIZACION_DE_LAS_BRECHAS_A_LOS_ALGORITMOS_04.pdf. Fecha de consulta: 08/09/2023.

UNESCO. (2019). *I'd blush if I could. Closeng gender divides en digital skills through education*. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=74. Fecha de consulta: 08/09/2023.

Wisner Glusko, D. C. (2022). Breves reflexiones sobre la importancia del Estado de Derecho en el desarrollo del marco legal sobre los sistemas de Inteligencia Artificial en la Unión Europea. En Llano Alonso, Fernando (dir.), Inteligencia Artificial y Filosofía del Derecho, Murcia: Laborum ediciones (pp. 529–545).

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., y Baeza-Yates, R. (2018). "FA IR: A fair top-k rankeng algorithm*". International Conference on Enformation and Knowledge Management, Proceedengs*, Part F1318, 1569–1578. https://doi.org/10.1145/3132847.3132938. Fecha de consulta: 08/09/2023.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., y Chang, K.-W. (2017). "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constrints", en *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark. Association for Computational Linguistics. (pp. 2979–2989) https://aclanthology.org/D17-1323/. Fecha de consulta: 08/09/2023.

Zhou, P., Shi, W., Zhao, J., Huang, K., Chen, M., Cotterell, R., Chang, K-W. (2019). "Examining Gender Bias in Languages with Grammatical Gender", en *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP*). Hong Kong, China. Association for Computational Linguistics. (pp. 5276–5284). https://aclanthology.org/D19-1531/. Fecha de consulta: 08/09/2023.

Zou, J. y Schiebinger, L. (2018). "AI can be sexist and racist – it's time to make it fair". *Nature.* https://www.nature.com/articles/d41586-018-05707-8. Fecha de consulta: 08/09/2023.