Sesgo de género (en IA)*
Gender bias (in AI)

María Pérez-Ugena Coromina
Universidad Rey Juan Carlos
ORCID ID 0000-0002-2724-6882
maria.perezugena@urjc.es

Cita recomendada:

Pérez-Ugena Coromina, M. (2024). Sesgo de género (en IA). *Eunomía. Revista en Cultura de la Legalidad*, 26, pp. 311-330

DOI: https://doi.org/10.20318/eunomia.2024.8515

Recibido / received: 03/12/2023
Aceptado / accepted: 18/01/2024

Resumen
Ante la emergencia de sistemas de inteligencia artificial en múltiples esferas, resulta imprescindible la incorporación de la perspectiva de género para impedir la perpetuación y fortalecimiento de estereotipos y conductas discriminatorias mediante la presencia de sesgos en los algoritmos.
La IA demanda una intervención consciente en su desarrollo y en la formulación de algoritmos, así como el cumplimiento de principios clave, como la transparencia y la responsabilidad, vitales para eludir el sesgo de género. El sesgo se demuestra mediante varios estudios y conlleva discriminación. Tiene su origen, generalmente, en la subrepresentación de grupos y se suele manifestar de forma indirecta. Se incluye un análisis del marco ético y regulatorio aplicable a los sistemas de inteligencia artificial en lo que resulta aplicable a la cuestión de género

Palabras clave
Inteligencia artificial; discriminación; igualdad; género.

Abstract
In the face of the emergence of artificial intelligence systems in multiple spheres, the incorporation of a gender perspective is essential to prevent the perpetuation and strengthening of stereotypes and discriminatory behaviors through biases in algorithms.
AI demands a conscious intervention in its development and algorithm formulation, along with adherence to key principles such as transparency and responsibility, vital to avoiding gender bias. Bias is demonstrated through various studies and entails discrimination. It generally originates from the underrepresentation of groups and tends to manifest indirectly. An analysis

* Profesora Titular de Derecho Constitucional de la Universidad Rey Juan Carlos. Forma parte del «Proyecto de Generación de Conocimiento» del Programa Estatal para Impulsar la Investigación Científico-Técnica y su Transferencia del Plan Estatal de Investigación Científica, Técnica y de Innovación 2021-2023, (2022/2026) PID2021-127122NB-I00. Miembro de grupo de investigación en Feminismo Género, de la Universidad Rey Juan Carlos.

of the regulatory framework applicable to artificial intelligence systems is included, as far as it is relevant to the gender issue.

Keywords
Artificial intelligent; equality; discrimination; gender.

SUMARIO. 1. Introducción. 2. Estereotipos de género e inteligencia artificial. 3. Efectos del uso de algoritmos sesgados y discriminación. 4. Marco ético y regulatorio aplicable. 5. Conclusiones.

1. Introducción

La incorporación de una perspectiva de género en la inteligencia artificial (IA) es esencial para prevenir la codificación inadvertida de estereotipos y sesgos de género en sus sistemas. Estos sesgos, profundamente arraigados en los datos generados por humanos, tienen el potencial de perpetuar y exacerbar la discriminación estructural preexistente. Los estereotipos de género, emergentes de procesos cognitivos, culturales y emocionales complejos, establecen expectativas y creencias sobre los roles «apropiados» para cada género. Desde una perspectiva interseccional, incluye también otros tipos de discriminación a los que está asociada de forma intrínseca la igualdad de género[1]. Si estos estereotipos se reflejan en la IA, no solo perpetúan la discriminación, sino que también la normalizan dentro de la estructura social.

La desigual valoración de los roles de género, con una tendencia a subordinar lo «femenino» a lo «masculino», se evidencia en la IA y demanda una intervención consciente en todas las fases de su ciclo de vida. Este trabajo recoge referencias a estudios con el propósito de demostrar que la IA puede discriminar en función del género, con dificultades particulares en el reconocimiento de voz de mujeres y en la detección facial de mujeres, especialmente de color, de manera similar a como lo han hecho históricamente los medios de comunicación. Para mitigar estos problemas, es crucial una selección de datos representativa y una revisión continua de los prejuicios en los algoritmos.

Este trabajo analiza la subrepresentación de las mujeres en el diseño y desarrollo de productos y servicios de IA, así como el uso de conjuntos de datos con sesgo de género, identificándolos como causas principales de sesgo discriminatorio. Persigue como objetivo plantear prácticas para limitar dicho sesgo.

El estudio del algoritmo, como componente central del sistema de IA, es fundamental para entender cómo controlar sus efectos en las posibles formas de discriminación de género que puede generar, ya sea de carácter estructural, directa, indirecta o en términos de igualdad formal o sustantiva.

En una etapa temprana de regulación de la IA, donde se están definiendo los límites legales y la necesidad de establecer controles, la aplicación de la perspectiva de género y la adopción de principios como la transparencia y la responsabilidad son cruciales. Este trabajo analiza la regulación actual de los sistemas de IA, incluyendo la específica de género, con el objetivo de comprender las normas y principios que inciden en la cuestión de género en el ámbito de la IA, haciendo referencia a los

[1] Vid. el «Informe preliminar con perspectiva de interseccionalidad sobre sesgos de género en la inteligencia artificial» (2023) publicado por el Instituto de la Mujer, p. 2. https://www.inmujeres.gob.es/areasTematicas/SocInfo/Estudios/docs/Informe_Sesgos_Genero_IA.pdf

acuerdos internacionales y a los sistemas de autorregulación de las grandes compañías tecnológicas.

2. Estereotipos de género e inteligencia artificial

Los estereotipos de género en la inteligencia artificial (IA) son un reflejo de los sesgos existentes en la sociedad[2]. Estos sesgos pueden ser introducidos en los sistemas de IA a través de los datos con los que se alimentan, que al estar creados por personas contienen su manera de entender el mundo. La IA puede reproducir y reforzar estereotipos de género y normas sociales discriminatorias si no se desarrolla y aplica con una perspectiva de género[3].

La perpetuación de estereotipos de género es, en efecto, una consecuencia directa de la de la propia organización social. Estos estereotipos no son meras etiquetas superficiales; son el resultado de procesos complejos y multifacéticos que incluyen aspectos cognitivos, culturales y emocionales. Estos procesos consolidan un conjunto de expectativas y creencias sobre las características y comportamientos «apropiados» para hombres y mujeres. Tales estereotipos no solo refuerzan la discriminación de género, sino que también se asumen como parte «natural» de la estructura social, lo que dificulta su erradicación.

Debe además tenerse en cuenta que las mujeres no son una minoría ni un grupo social entre otros, sino la mita de la humanidad y la mitad que ha estado excluida de lo público, la ciencia, la técnica y el rumbo de estas durante milenios. Lo que ha generado el negativo desequilibrio que afecta a todo lo humano y, también, referido a este caso, al desarrollo de la IA.

Los roles y funciones que la sociedad y la cultura asignan a lo que se considera «femenino» suelen estar subordinados y reciben menos reconocimiento que aquellos asociados con lo «masculino». Esta disparidad en la valoración social de los roles de género es evidente y perjudicial. Para contrarrestar esta tendencia, especialmente con la creciente influencia de los sistemas de inteligencia artificial (IA), es imperativo intervenir en todas las fases del ciclo de vida de estos sistemas. Esto implica desde la concepción y diseño hasta el desarrollo, implementación y monitoreo continuo de la IA[4].

Llevado al ámbito de la IA, el estereotipo o comportamiento discriminatorio tiene un significado más estrecho y hace referencia a la existencia de cualquier diferencia estadísticamente significativa en las acciones agregadas de robots en función de la raza (por ejemplo, negra vs. asiática, género –por ejemplo, mujer vs. Hombre–, o la intersección de ambas categorías -por ejemplo, mujer negra vs. hombre asiático). (Hundt, Agnew, Zeng, Kacianka y Gombolay, 2022)[5]. Varios estudios han

[2] Vid Naciones Unidas sobre igualdad de género y empoderamiento de las mujeres. https://www.un.org/womenwatch/daw/beijing/platform/

[3] La definición que el Comité Preparatorio de Pekín propuso de «género» es la siguiente: «Género se refiere a los roles y responsabilidades de la mujer y del hombre que son determinados socialmente. El género se relaciona con la forma en que se nos percibe y se espera que pensemos y actuemos como mujeres y hombres, por la forma en que la sociedad está organizada, no por nuestras diferencias biológicas». El género se refiere, por tanto, a las relaciones entre mujeres y hombres basadas en roles definidos socialmente, que se asignan a uno u otro sexo.

[4] En este sentido se ha pronunciado la UNESCO Recomendación sobre la ética de la inteligencia artificial, adoptada el 23 de noviembre de 2021. https://www.unesco.org/es/artificial-intelligence/recommendation-ethics

[5] Vid también: https://incidentdatabase.ai/es/. La base de datos de incidentes de IA está dedicada a indexar el historial colectivo de daños o casi daños realizados en el mundo real por el despliegue de sistemas de inteligencia artificial.

constatado que la IA, especialmente en métodos de aprendizaje automático, sistemas de procesamiento del lenguaje natural, o ambos, en el caso de modelos grandes de imágenes y leyendas como *Open AI CLIP,* estereotipa por razón de raza, género, fisonomía científicamente desacreditada más que los propios humanos (Hundt, Agnew, Zeng, Kacianka y Gombolay, 2022).

La introducción de sesgos puede acontecer de dos maneras fundamentales. En primer término, mediante la selección de datos que no representan fielmente la realidad, un fenómeno conocido como sesgo de muestreo. Un ejemplo concreto sería la preferencia por utilizar más imágenes de hombres que de mujeres, generando así una distorsión en la representación de la diversidad.

En segundo lugar, el sesgo puede emerger al reflejar prejuicios ya existentes en los datos de entrenamiento del algoritmo. Por ejemplo, la utilización de información histórica sobre decisiones de contratación, que favorezcan a los hombres en detrimento de las mujeres, llevaría al algoritmo a aprender y perpetuar esa discriminación.

Se han realizado algunos estudios en los que se observan mayores dificultades de reconocer a mujeres y personas de color. Se han estudiado los casos que incluyen la incapacidad para reconocer a personas con tonos de piel oscuros. Se producen sesgos en la detección facial, donde los hombres con los tonos de piel más claros son detectados con mayor precisión, las mujeres con el tono de piel más claro menos, y las mujeres con los tonos de piel más oscuros con una precisión dramáticamente inferior (Buolamwini y Timnit, 2018).

Los sistemas de reconocimiento de voz contienen también sesgos. En concreto plantean problemas con las voces agudas. La razón subyacente puede ser que las bases de datos tienen muchos datos sobre hombres y menos datos sobre voces femeninas y minoritarias, por lo que a menudo funcionan peor para las mujeres[6].

Los asistentes personales virtuales (VPAs), como Alexa y Siri son un ejemplo evidente de cómo la IA puede reforzar los estereotipos de género. Estos son sistemas de IA que pueden interactuar con las personas mediante el lenguaje natural, y que pueden ofrecer información, servicios o entretenimiento. Muchos de estos VPAs tienen una voz femenina por defecto, y se les asignan nombres y funciones que se asocian tradicionalmente con las mujeres, como cuidar, asistir y encargarse del hogar[7].

El uso predominante de voces masculinas en los medios de comunicación ha venido utilizándose mayoritariamente, especialmente en publicidad, por resultar de mayor autoridad y otorgar mayor credibilidad. Como destaca un informe de la UNESCO de 2019[8], no es coincidencia que los servicios de asistentes personales virtuales como Siri, Alexa y Cortana tengan nombres femeninos y voces femeninas predeterminadas. Las empresas que diseñan estos servicios refuerzan una realidad social en la que la mayoría de las personas que prestan servicios secretariales o de asistencia personal en los sectores públicos y privados son mujeres.

---

[6] Vid. *Impacto del sesgo de género y raza en la ia.* https://ciberseguridad.com/guias/nuevas-tecnologias/inteligencia-artificial/sesgo-genero-raza/

[7] Biblioteca UNESCO. *"I'd blush if I could: closing gender divides in digital skills through education*". https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

[8] *Ibid.*

Además, las diferencias de género en publicidad también se manifiestan en el tipo de productos que anuncian hombres y mujeres. Así, los anuncios que pueden situarse en el ámbito doméstico se hacen fundamentalmente por mujeres. También ha sido analizada la posición que ocupan hombres y mujeres, que es distinta. Los hombres ocupan estatus superiores y mejores empleos. Destacan una ausencia de mujeres abogadas, doctoras, ejecutivas, científicas, ingenieras, atletas y similares. Por último, y también en relación con la voz, la utilización de voces masculina-femenina ponen de manifiesto la relación entre femenino y doméstico. Así, las voces masculinas se utilizan en todo tipo de productos y las femeninas básicamente en productos comerciales de carácter doméstico (Coltrane y Messineo, 2000).

La utilización de estos asistentes personales transmite el mensaje de que las mujeres son serviciales, sumisas y dependientes, y que su trabajo es menos valorado y reconocido que el de los hombres. Además, algunos de estos *VPAs* pueden responder de forma pasiva o complaciente ante insultos o comentarios sexistas, lo que puede fomentar la violencia y el acoso contra las mujeres. Esto recuerda la extensa literatura existente sobre la forma en que las mujeres han aparecido tradicionalmente en los medios de comunicación sin que representen una imagen acorde con la realidad. Al contrario, las mujeres aparecen generalmente poseedoras de un poder más reducido y débiles, además de que se les sitúa en una posición de sumisión y generalmente hablando a niños o animales. En un sentido general las mujeres no suelen aparecen representando figuras que incluyan autoridad o como expertas o personas que hablen a un público (Lovdal, 1989). Hay un claro desequilibrio entre la aparición en los medios de hombres y mujeres. De tal forma que no solo la diferencia es cualitativa, también cuantitativa. Sólo una de cada tres personas que aparecen en televisión son mujeres[9].

Además, los roles sexuales están basados en un doble estándar en el que con demasiada frecuencia las mujeres aparecen como objetos sexuales cuyo valor viene dado exclusivamente por su apariencia física y su atractivo sexual (Janna, 2007)[10]. De forma parecida ocurre con las imágenes que surgen de sistemas de IA, en las que se refuerzan estereotipos por influencia del sesgo. Se ha demostrado que también afecta a conjuntos de datos ampliamente utilizados en aprendizaje automático, como *ImageNet y Open Images*. Los sesgos de representación en estos conjuntos son claros, y se ha abogado por la inclusión de diversidad geográfica como medida de mitigación. En el ámbito del Procesamiento del Lenguaje Natural (NLP), se identifican sesgos representacionales en bases de conocimiento utilizadas en diversas aplicaciones, como, por ejemplo, la asociación estereotipada de masculinidad y feminidad con las ciencias y las artes respectivamente (Mehrabi, Morstatter, Saxena, Lerman y Galstyan, 2019).

La causa principal del sesgo en género es, por tanto, la subrepresentación de las mujeres en el diseño y desarrollo de productos y servicios de IA, así como el uso de conjuntos de datos con sesgo de género. Además, es necesario considerar no solo géneros, sino también subdivisiones por raza para lograr una representación más precisa (Mehrabi, Morstatter, Saxena, Lerman y Galstyan, 2019) La inclusión de una amplia representación fenotípica y demográfica en los conjuntos de datos de rostros es fundamental para evitar el sesgo (Buolamwini y Timnit, 2018, pp. 1-15).

[9] https://waccglobal.org/our-work/global-media-monitoring-project-gmmp/. Se trata de la más extensa investigación global que se ha hecho en cuestión de género con relación a noticias en los medios de comunicación.

[10] En este estudio se analizaban los códigos con contenido sexual desde una perspectiva feminista elaborado sobre 25 programas de televisión más vistos por adolescentes en EE. UU.

La transparencia y la responsabilidad se han destacado como principios que se deben cumplir para evitar el sesgo, así como rendición de cuentas y la fiabilidad de los algoritmos que se utilicen[11].

La transparencia no es un concepto estático y su contenido evoluciona según la finalidad y el contexto en el que se aplique (Cotino Hueso y Castellanos Claramunt, 2022). Se logra al proporcionar detalles sobre la composición de los conjuntos de datos de entrenamiento y referencia.

La responsabilidad implica que los actores de la IA asuman las consecuencias de sus acciones y decisiones, y que puedan ser sometidos a sanciones o reparaciones en caso de que causen daños o violen los derechos. Se manifiesta al informar sobre el rendimiento algorítmico en subgrupos específicos y al esforzarse por mitigar las disparidades.

La rendición de cuentas obliga a que los actores de la IA puedan demostrar y justificar que sus acciones y decisiones son conformes con las normas y los principios aplicables, y que puedan ser objeto de control y supervisión por parte de las autoridades competentes. La fiabilidad, por último, supone que los sistemas de IA sean seguros, robustos y resistentes a los errores, las manipulaciones y los ataques. Estos principios son esenciales para avanzar hacia una inteligencia artificial que sea justa y confiable para todos los usuarios (Buolamwini y Timnit, 2018, pp. 1-15).

En definitiva, puede observarse que se da continuidad a los estereotipos que ya se difundían en los medios de comunicación y que venían a perpetuar una desigualdad en la sociedad al reproducir y dar fuerza a los patrones existentes como consecuencia de una distribución de roles que ni siquiera es la importante en muchos casos en la sociedad y que viene reflejada por los medios (Matthes, Prieler, y Adam, 2016).

Esto no es solo un problema de la IA actual, sino también de la IA futura. Las múltiples fuentes de sesgo de género, así como las particularidades de cada tipo de algoritmo y conjunto de datos, hacen que eliminar el sesgo sea una cuestión compleja. Pero, si no se cambia el *statu quo*, los estereotipos de género de hoy pueden influir en las tecnologías de mañana, y perpetuar la desigualdad y la discriminación entre mujeres y hombres. Por eso, es necesario promover una IA que respete la diversidad, la inclusión y la equidad de género, y que contribuya al empoderamiento y al progreso de las mujeres.

Ya que el riesgo de amplificar físicamente estereotipos negativos en general; y que corregir simplemente las disparidades será insuficiente para la complejidad y escala del problema deben detenerse, reformularse o incluso reducirse, según sea apropiado, los métodos de aprendizaje de robots que manifiesten estereotipos u otros resultados perjudiciales, hasta que se demuestre que los resultados sean seguros, efectivos y justos. (Hundt, Agnew, Zeng, Kacianka y Gombolay, 2022). Para mitigar la reproducción de estereotipos de género en la IA se propone implementar prácticas como auditorías de sesgo, con revisiones periódicas de los conjuntos de datos y algoritmos para identificar y corregir sesgos de género. También la educación y sensibilización son aspectos fundamentales, para fomentar la conciencia sobre la importancia de la equidad de género en la tecnología entre los profesionales de la IA y el público en general. El diseño de los sistemas debe ser inclusivo, de forma que se

[11] Recomendación sobre la ética de la inteligencia artificial, adoptada el 23 de noviembre de 2021 por la UNESCO. https://www.unesco.org/es/artificial-intelligence/recommendation-ethics

asegure que los productos de IA reflejen la diversidad de usuarios y no perpetúen roles de género obsoletos o dañinos.

Por último, son fundamentales las regulaciones y políticas que establezcan marcos legales y éticos que guíen el desarrollo y uso responsable de la IA en relación con la equidad de género.

3. Efectos del uso de algoritmos sesgados y discriminación

Como se ha señalado, la inteligencia artificial reproduce estereotipos mediante la inclusión de datos sesgados que devienen de actitudes discriminatorias. El algoritmo es la pieza clave de todo el engranaje que, al expresarse en un lenguaje de programación, permite que se realicen actividades registradas y procesadas por sistemas inteligentes. Las Inteligencias artificiales utilizan algoritmos para procesar grandes cantidades de datos y tomar decisiones basadas en patrones y reglas establecidas a través del aprendizaje automático.

Puede definirse como un «conjunto de instrucciones o reglas definidas y no-ambiguas, ordenadas y finitas que permite, típicamente, contestar una pregunta, tomar una decisión, solucionar un problema, realizar un cómputo, procesar datos o llevar a cabo alguna tarea». Estos procedimientos computacionales toman uno o varios valores de entrada y generan uno o varios valores de salida, por lo tanto, son instrumentos que no intentan establecer un vínculo causal entre una variable específica y su efecto, sino que producen un resultado[12].

Desde la perspectiva de su ámbito de acción, hay infinidad de tipos de algoritmos, los que interesan desde la óptica del Derecho Público son aquellos que afectan a sectores públicos o de interés social; aquellos que guardan relación con el acceso a ciertas prestaciones que se relacionan con los derechos fundamentales o con bienes constitucionalmente protegidos. Es decir, cuestiones como el acceso a la educación o al trabajo, otras que conllevan discriminación basada en algún posible aspecto derivado del artículo 14 de la Constitución, o que se implementan en el ámbito judicial.

El algoritmo va a depender de cómo lo hayan previsto sus diseñadores, desarrolladores y operadores. En el trabajo de buscadores se accede a la información según ésta se haya categorizado. De forma que los resultados se muestran no solo según las peticiones de los usuarios sino según se hayan categorizado.

Básicamente, el problema que se plantea con el uso de los algoritmos, especialmente aquellos de aprendizaje automático, a los que se ha definido como «cajas negras» es la dificultad de que las decisiones derivadas puedan ser comprendidas o controladas, al convertirse en conjuntos opacos de códigos informáticos y datos (Cotino Hueso y Castellanos Claramunt, 2022). Por eso, resulta fundamental que los algoritmos se diseñen, desarrollen y utilicen de forma que sean, no solo adecuados desde el punto de vista jurídico, sino también deben ser fáciles de supervisar y corregir, controlando cualquier sesgo. La transparencia se convierte en un aspecto fundamental para prevenir, identificar y mitigar posibles sesgos[13].

[12] *Guía de auditoría algorítmica. Éticas-consulting.* https://www.eticasconsulting.com/eticas-consulting-guia-de-auditoria-algoritmica-para-desarrollar-algoritmos-justos-y-eficaces/
[13] *Ibid.*

Se puede precisar que el objetivo de análisis a la hora de estudiar cómo se afecta a los derechos fundamentales, no son los algoritmos en sí mismos, sino los procesos de toma de decisiones en torno a los algoritmos.

Se plantean algunas cuestiones en el desarrollo del derecho digital relativas a cómo se integran los algoritmos en el panorama legal. O si se requiere desarrollar nuevas categorías legales para abordar los efectos de los algoritmos en la sociedad[14].

Respecto de la naturaleza jurídica del algoritmo, lo determinante debería ser el ámbito específico en el que se utiliza el algoritmo. De tal forma, que su régimen jurídico no se haga depender de la naturaleza de los procesos técnicos que realizan ni del código fuente en sí mismo considerado.

En el ámbito concreto de la Administración Pública tampoco existe unanimidad acerca del carácter o naturaleza jurídica de los algoritmos (Gutiérrez David, 2021, pp. 55-56 y Vestri, 2021) aunque puede aceptarse el planteamiento de que, si materialmente realizan las mismas funciones que los reglamentos, deben estar sometidos a las similares garantías (Martín Delgado, 2009 y Boix Palop, 2020).

En lo que si hay consenso es en la necesidad de que se regule, puesto que el marco jurídico existente en la actualidad no es suficiente ni permite que las decisiones tomadas sobre algoritmos sean trazables o explicables.

La introducción de discriminación potencial en un sistema puede surgir por motivos étnicos, de estatus económico, género, edad, demografía, religión u otros factores, con la tendencia a perjudicar a minorías o grupos insuficientemente representados en los datos que se emplean como referentes en el aprendizaje de los sistemas de IA. Estos derivan del hecho de que los sistemas de aprendizaje automático se diseñan y se entrenan con datos, que, en la mayoría de los casos, fueron seleccionados por humanos (Mehrabi, Morstatter, Saxena, Lerman, y Galstyan, 2019).

El efecto discriminatorio se produce como consecuencia de un trato injustificado a ciertos individuos o grupos de individuos a favor de otros, por motivos que no son razonables o inapropiados. Lo que resulta especialmente preocupante por el hecho de que los algoritmos, a diferencia de los humanos, no tienen capacidad para contrarrestar conscientemente los sesgos que hayan podido incorporar, tanto de forma consciente como inconsciente, sus desarrolladores.

La discriminación se manifiesta en sus distintas formas o significados, tanto referida a la continuidad de la discriminación estructural, y sus efectos, como en la discriminación tanto directa como indirecta y en su aspecto formal y real.

Respecto de la discriminación estructural, es evidente que la IA, sino se controla el sesgo que, de hecho, existe, va a producir una mayor desigualdad estructural entre hombres y mujeres. Se han señalado distintos ámbitos o sectores en los que se ejemplifican estas situaciones. Salud, empleo, movilidad, arquitectura y trabajo policial[15].

También el sesgo implica un riesgo claro de discriminación, en el sentido del artículo 14 de la Constitución, referida a la igualdad puramente formal frente a la discriminación por razón de nacimiento, raza, sexo, religión, opinión o cualquier otra

[14] *Ibid*.
[15] Vid informe preliminar con perspectiva interseccional sobre sesgos de género en la inteligencia artificial https://www.inmujeres.gob.es/areasTematicas/SocInfo/Estudios/docs/Informe_Sesgos_Genero_IA.pdf

condición o circunstancia personal o social. Es preciso establecer esta relación, puesto que conlleva un refuerzo en el sistema de garantías, al tratarse no solo de un principio, el de igualdad, que tiene carácter inspirador, sino también de un derecho, que es, además, un derecho fundamental. Y se garantiza, por tanto, de acuerdo con lo que el ordenamiento jurídico español prevé para el sistema de derechos fundamentales.

Y, por último, la igualdad real y efectiva a que hace referencia el artículo 9.2 de la Constitución española exige la puesta en marcha de acciones por parte del poder público para mitigar la desigualdad de hecho. Los sistemas IA tienen la capacidad de situar a las mujeres en una posición de desventaja que debe necesariamente suplirse con el instrumento, recogido en el artículo 9.2 de la Constitución, de acuerdo con el cual, «corresponde a los poderes públicos promover las condiciones para que la libertad y la igualdad del individuo y de los grupos en que se integra sean reales y efectivas; remover los obstáculos que impidan o dificulten su plenitud y facilitar la participación de todos los ciudadanos en la vida política, económica, cultural y social». De lo anterior derivan las posibles tomas de acciones positivas que puedan contrarrestar los efectos de los sistemas IA en el aumento de la brecha de género.

Se pueden producir situaciones de discriminación directa, o indirecta como efecto derivado de la creación, desarrollo o implementación de los sistemas IA.

En el primer caso, la discriminación puede ser resultante de la acción, o norma, que se dirige a tratar de forma diferenciada y desfavorable a una persona o a determinados grupos o colectivos[16]. Es discriminación directa la que se basa, por ejemplo, en el sexo o en alguna característica que se relacione con ésta.

También pueden plantearse supuestos de discriminación indirecta, resultado de acciones que no tienen, o no parecen tener, por objeto un trato discriminatorio, pero de su aplicación práctica resulta, de hecho, tal discriminación[17]. Es decir, se aplica un criterio que aparentemente es neutral, pero provoca efectos desproporcionadamente desiguales para uno de los sexos (Serra, 2004). La aparición de sesgos en la IA es mayoritariamente resultado de formas de discriminación indirecta. La discriminación indirecta no guarda relación con la intencionalidad, es decir, la medida de que se trate no pretende discriminar, pero el efecto que produce en su aplicación, el resultado, es la discriminación de las mujeres.

La demostración o prueba de que existe discriminación, en este caso, exige una comparación, a través de una constatación estadística, que ponga de manifiesto el grado en que una decisión afecta a los distintos grupos, que no tienen que ser homogéneos, sino que se toma en cuenta la composición mayoritaria. Es un ejemplo

[16] De acuerdo con la Directiva 2002/73/CE del Parlamento Europeo y del Consejo, es discriminación directa «*aquella situación en que una persona es, haya sido, o pueda ser tratada de manera menos favorable que en otra situación comparable por razón de sexo*».

[17] Es ilustrativo el caso recogido en la STC 145/1991, que se refiere al supuesto de limpiadoras de un hospital público que cobraban un salario inferior a los peones. La doctrina del Tribunal Constitucional a partir de esta Sentencia prohíbe la desigual valoración de trabajos equivalente cuando este tratamiento diferenciado atienda al sexo de quienes trabajan. Esto supone que, teniendo en cuenta que la mayoría de las mujeres son las que ocupan, en ese caso concreto, el puesto de limpiadoras y la mayoría de los hombres son los que ocupan el puesto de peón, con esta medida de dar un salario inferior a limpiadoras se está perjudicando, de hecho, a las mujeres. Aún, cuando el perjuicio se produzca de manera indirecta. En estos supuestos el TC ha entendido que la diferencia pasa a ser sospechosa a menos que se justifique que no se funda en el sexo, sino en las características del trabajo. De ahí que la igualdad de retribuciones no solo deba ser la misma para un mismo trabajo sino también para un trabajo distinto que tenga igual valor (en SSTC 198/1996; 240/1999).

claro de supuestos de discriminación indirecta a través de IA, la forma de reclutamiento de Amazon, que se basaba en un algoritmo de aprendizaje automático. Ésta mostró una preferencia por los candidatos masculinos y penalizó a las mujeres que aspiraban a puestos técnicos. Lo que se debió a que el algoritmo aprendió de los datos históricos de contratación de la empresa, en los que predominaban los hombres. El algoritmo también asignó una puntuación más baja a los currículos que contenían la palabra «mujer» o el nombre de universidades femeninas. A pesar de los esfuerzos por hacer el algoritmo neutral a estos términos, no se pudo garantizar que no desarrollaría otras formas de discriminación. Amazon intentó corregir el sesgo del algoritmo, pero finalmente decidió abandonar el proyecto por falta de confianza en su neutralidad (Kullmann, 2018, p. 20)[18].

Este caso ilustra cómo los sesgos en los datos de entrenamiento pueden llevar a la IA a perpetuar discriminaciones existentes, lo que resalta la importancia de utilizar datos diversificados y realizar auditorías de sesgo para desarrollar sistemas de IA justos y equitativos.

Es crucial destacar que estos sesgos no se limitan a simples errores técnicos; sus implicaciones son significativas. Se ha observado que tienden a repetirse y a perpetuarse, fortaleciendo dinámicas de dominación, privilegio y discriminación. En última instancia, esto amplifica el riesgo de que las desigualdades existentes se vean acentuadas y consolidadas a través de la automatización y el aprendizaje de las máquinas.

Por último, hay otro tipo de sesgo que puede ser aún más preocupante, y es el sesgo en la interpretación de los resultados de la inteligencia artificial. Este sesgo humano se manifiesta cuando aceptamos de manera acrítica los resultados de un sistema de inteligencia artificial como verdaderos e inamovibles, adoptando un «principio de autoridad» basado en las expectativas generadas por dichos sistemas[19].

En otras palabras, este sesgo implica confiar ciegamente en los resultados de la IA sin cuestionar ni analizar de manera crítica su validez. Esto puede llevar a decisiones erróneas o injustas, ya que la interpretación sesgada de los resultados puede estar influenciada por prejuicios o expectativas no fundamentadas. En consecuencia, es esencial abordar tanto los sesgos inherentes en los datos y algoritmos como el sesgo humano en la interpretación de los resultados para garantizar una toma de decisiones más objetiva y equitativa[20].

4. Marco ético y regulatorio aplicable

En primer lugar, es preciso distinguir entre la normativa o la regulación aplicable a la inteligencia artificial de una manera general y hacer una reflexión al marco general en el que se aplica la cuestión de género en este ámbito. En segundo orden y de manera específica, se tratan los documentos más importantes que ya se han aprobado respecto de la inteligencia artificial y sus efectos en el ámbito del género.

De manera general, resulta necesario señalar que la inteligencia artificial, pese a que pudiera parecer lo contrario en una primera impresión, es un sector hiper-regulado. Le afectan normas de distintos tipos, naturaleza y niveles. Lo que se precisa es crear marcos generales en los que la ética juega un papel determinante

[18] Sobre el algoritmo de Amazon con efectos discriminatorios, vid: https://www.bbc.com/mundo/noticias-45823470

[19] Vid informe sobre Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción https://www.aepd.es/documento/adecuacion-rgpd-ia.pdf

[20] *Ibid*.

(Hernández Peña, 2022) En este sentido, a nivel global, La UE ha desarrollado una estrategia digital, que se fundamenta en una serie de principios[21] y pretende una transformación digital protegida, segura y sostenible que sitúe a las personas en el centro, en consonancia con los valores y los derechos fundamentales de la UE.

Con ese propósito, en febrero de 2020, la Comisión publicó el Libro Blanco sobre la inteligencia artificial: un enfoque europeo orientado a la excelencia y la confianza[22]. En el Libro Blanco se definen las opciones existentes para alcanzar el doble objetivo de promover la adopción de la IA y de abordar los riesgos vinculados a determinados usos de esta nueva tecnología.

Recientemente, la Declaración Europea sobre los Derechos y Principios Digitales para la Década Digital expresa el compromiso de la Unión Europea con una transformación digital centrada en las personas, en consonancia con los valores fundamentales de la UE y los derechos fundamentales. Fue firmada por los presidentes del Parlamento Europeo, el Consejo y la Comisión el 15 de diciembre de 2022 [23].

La Declaración contiene seis puntos clave, que abordan aspectos como la soberanía digital abierta, el respeto de los derechos fundamentales, la inclusión, la accesibilidad, la igualdad y la no discriminación, la participación en el espacio público digital, la seguridad, la protección y el empoderamiento, y la sostenibilidad en el entorno digital. La Declaración pretende ser un marco de referencia para el desarrollo de una IA ética y responsable, que garantice la protección de los derechos humanos y el bienestar de la sociedad.

En el marco de esa política digital, en abril de 2021, la Comisión Europea presentó una propuesta para establecer un marco normativo de la Unión Europea sobre inteligencia artificial (IA). Esta iniciativa, conocida como el proyecto de ley de IA, tiene como objetivo principal regular de manera horizontal el campo de la IA. El marco legal propuesto se centra en la utilización específica de los sistemas de IA y los riesgos asociados. Con este marco se pretende recoger un enfoque europeo coordinado sobre las implicaciones éticas y humanas de la IA.

Recientemente, en diciembre de 2023, el Consejo y el Parlamento Europeo han llegado a un acuerdo sobre las primeras normas del mundo en materia de inteligencia artificial.

Los puntos principales del acuerdo incluyen:

En primer lugar, normas sobre modelos de IA de uso general de gran impacto que pueden causar un riesgo sistémico en el futuro, los sistemas de IA de alto riesgo, así como un sistema revisado de gobernanza con algunas competencias de ejecución a escala de la UE.

---

[21] La Declaración, presentada por la Comisión en enero de 2022 señala una serie de principios relacionados con la transformación digital. https://wayback.archive-it.org/12090/*/https://ec.europa.eu/digital-single-market/en/news/digital-single-market-strategy-europe-com2015-192-final

[22] Comisión Europea, *Libro Blanco sobre la inteligencia artificial: un enfoque europeo orientado a la excelencia y la confianza*, COM (2020) 65 final, 2020. https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52020DC0065

[23] Declaración Europea sobre los Derechos y Principios Digitales para la Década Digital (2023/C 23/01). https://eurlex.europa.eu/legalcontent/ES/TXT/PDF/?uri=CELEX:32023C0123(01)&qid=1684248405089

Se da lugar a una ampliación de la lista de prohibiciones, pero con la posibilidad de utilizar la identificación biométrica remota por parte de las autoridades policiales en espacios públicos, con sujeción a salvaguardias.

Una mejor protección de los derechos mediante la obligación de que los implementadores de sistemas de IA de alto riesgo lleven a cabo una evaluación del impacto en los derechos fundamentales antes de poner en marcha un sistema de IA.

El objetivo, según el acuerdo, es garantizar que los sistemas de IA comercializados en el mercado europeo y utilizados en la UE sean seguros y respeten los derechos fundamentales y los valores de la UE. Esta propuesta histórica también pretende estimular la inversión y la innovación en IA en Europa. Como la primera propuesta legislativa de este tipo en el mundo, puede establecer un estándar global para la regulación de la IA en otras jurisdicciones, al igual que el RGPD (Reglamento General de Protección de Datos), promoviendo así el enfoque europeo de la regulación tecnológica en el escenario mundial.

La Unión Europea ha adoptado un enfoque humanista y centrado en la persona para regular la inteligencia artificial, especialmente en lo que respecta a la protección de los derechos fundamentales. Este enfoque es fundamental frente a los riesgos de que se produzcan arbitrariedades, sesgos, discriminaciones y otras vulneraciones de los Derechos fundamentales. Por ello, el regulador europeo ha iniciado desde el año 2018 un proceso de consulta y trabajo con diversos actores para establecer un marco normativo adecuado.

La propuesta del regulador europeo se basa en un enfoque basado en riesgos, que tiene en cuenta la probabilidad y el impacto de los posibles daños derivados de la inteligencia artificial generativa. Según este enfoque, se distinguen cuatro niveles de riesgo: inaceptable, alto, limitado y mínimo (Hernández Peña, 2022).

En el nivel inaceptable, se prohíben los usos de sistemas IA que supongan una manipulación cognitiva o del comportamiento de las personas, especialmente de las más vulnerables, como los menores, por ejemplo, juguetes activados por voz que fomentan comportamientos peligrosos en los niños.

También se prohíben los usos que impliquen una puntuación social, clasificación de personas en función de su comportamiento, estatus socioeconómico o características personales; o una privación de derechos o servicios esenciales.

Por otro lado, se autorizaría el uso de una amplia gama de sistemas de IA de «alto riesgo», pero con la condición de cumplir con una serie de requisitos y obligaciones para poder acceder al mercado de la UE[24]. En este grupo estarían los usos de sistemas IA que tengan un impacto significativo en la vida o los derechos de las personas, que quedan sometidos a una serie de requisitos y garantías.

En el nivel limitado, se aplican medidas de transparencia y responsabilidad a los usos de sistemas IA que puedan afectar a la calidad o la seguridad de los productos o servicios. En el nivel mínimo, se establecen unas normas generales de buena práctica para los usos de los sistemas IA que no entrañen riesgos significativos.

Un caso particular que merece una atención especial es el de los sistemas biométricos, que pueden hacer uso de la IA para reconocer o identificar a las personas. Estos sistemas se consideran de alto riesgo cuando se aplican en tiempo

[24] Todos los sistemas de IA de alto riesgo serán evaluados antes de su comercialización y a lo largo de su ciclo de vida.

real y a distancia, salvo que se den ciertas condiciones excepcionales, como la prevención o investigación de delitos graves, la protección de la seguridad pública o la defensa nacional. En estos casos, se requiere una autorización previa y una supervisión adecuada.

Por otra parte, se consideran de riesgo limitado o mínimo el uso de inteligencia artificial que forme parte de productos o sistemas que ya están sujetos a normas de seguridad específicas, como los dispositivos móviles, el transporte, los dispositivos médicos, los ascensores o los juguetes. En estos casos, se permite la integración de la inteligencia artificial generativa como un componente más, siempre que se respeten las normas aplicables y se informe adecuadamente al usuario. Asimismo, se permite la identificación biométrica a partir de rasgos como las huellas dactilares o la firma, siempre que se cumplan los requisitos legales pertinentes.

Los sistemas IA también se permiten en el ámbito de la educación, el empleo, la selección de personal, el acceso a servicios públicos o la migración, siempre que se garantice el respeto a los derechos y libertades. Además, se establece que toda asistencia que tenga que ver con la interpretación o la aplicación de la ley debe estar sujeta a determinados procedimientos y garantías.

Para asegurar el cumplimiento de estos requisitos y garantías, el regulador europeo ha establecido un sistema de evaluación, registro, información, supervisión y control de la inteligencia artificial generativa. Antes de producir o comercializar cualquier producto o sistema que haga uso de la inteligencia artificial generativa, se debe realizar un análisis de riesgo *ex ante* o por diseño, que verifique que se respeta el ordenamiento jurídico y los principios éticos. Además, se debe registrar el producto o sistema en una base de datos pública, que permita su trazabilidad y seguimiento. Asimismo, se debe informar al usuario de forma clara y adecuada de que está utilizando un producto o sistema que incorpora inteligencia artificial generativa, y de los posibles riesgos o beneficios que ello conlleva. Por último, se debe realizar una evaluación continua y periódica de los riesgos y el impacto de la inteligencia artificial generativa, y se debe establecer un mecanismo de supervisión y control por parte de las autoridades competentes.

La UE pretende, con esta regulación, situarse como líder mundial en el desarrollo de sistemas IA seguros y éticos[25]. En diciembre de 2021, el Consejo de la UE acordó la posición general de los Estados miembros sobre esta propuesta. Posteriormente, el 14 de junio de 2023, el Parlamento Europeo emitió su posición al respecto.

Actualmente, se trabaja en las negociaciones para finalizar la nueva regulación integral de la IA, que implicarán enmiendas sustanciales a la propuesta inicial de la Comisión. Estas enmiendas incluirían la revisión de la definición de los sistemas de IA, la ampliación de la lista de sistemas de IA prohibidos y la imposición de obligaciones para la IA de propósito general y los modelos de IA generativa, como *ChatGPT*[26].

[25] Comisión Europea (2021a): "Generar confianza mediante el primer marco jurídico sobre la IA". En *Excelencia y confianza en la inteligencia artificial*. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_es#generar-confianza-mediante-el-primer-marco-jurdico-sobre-la-ia. Comisión Europea (2021b): Nuevas normas sobre la inteligencia artificial: preguntas y respuestas. https://ec.europa.eu/commission/presscorner/detail/es/QANDA_21_1683

[26] La IA generativa, como ChatGPT, tendría que cumplir requisitos de transparencia: 1) revelar que el contenido ha sido generado por IA; 2) diseñar el modelo para evitar que genere contenidos ilegales; 3) publicar resúmenes de los datos protegidos por derechos de autor utilizados para el entrenamiento

Las dificultades que plantea crear un marco regulador a esta materia exigen la búsqueda de criterios éticos en búsqueda de soluciones globales acordadas. El intento de crear un marco regulador de la IA es muy complejo, especialmente por el hecho de que los avances tecnológicos no son claramente previsibles y es preciso esperar a que se produzcan para que se plantee la necesidad de implementar normas jurídicas. Es decir, no hay una suficiente capacidad de anticipación en este sentido. De ahí que resulte imprescindible aplicar el Derecho que venga a proteger a la ciudadanía de una forma lo más abierta posible, en el sentido de lo más neutral.

El derecho va siempre detrás de la tecnología siendo muy difícil que se anticipe. El caso de los sistemas IA, al tratarse una tecnología disruptiva, es mucho más complicado suponer cuáles son los cambios tecnológicos que se van a llevar a cabo y cómo van a afectar al régimen de derechos. La ética juega, por estos motivos, un papel fundamental. También los códigos de autorregulación cumplen una función muy relevante (Hernández Peña, 2022).

En este sentido, la Recomendación sobre la ética de la inteligencia artificial, de la UNESCO adoptada el 23 de noviembre de 2021[27] representa un hito significativo en el ámbito de la ética de la inteligencia artificial, al convertirse en el primer marco normativo universal sobre esta cuestión, y que tiene como objetivo orientar a los Estados, las organizaciones internacionales, el sector privado, la sociedad civil y la comunidad científica en el desarrollo y el uso de la IA, de acuerdo con los valores universales de los derechos humanos, la dignidad humana y la diversidad cultural[28].

La Recomendación defiende ciertos valores y descansa en determinados principios, referidos a distintos ámbitos de actuación política. Aborda la ética de la IA como una:

> Reflexión normativa sistemática, basada en un marco integral, global, multicultural y evolutivo de valores, principios y acciones interdependientes, que puede guiar a las sociedades a la hora de afrontar de manera responsable los efectos conocidos y desconocidos de las tecnologías de la IA en los seres humanos, las sociedades y el medio ambiente y los ecosistemas, y les ofrece una base para aceptar o rechazar las tecnologías de la IA.

El enfoque central de esta declaración radica en la promoción y salvaguardia de los derechos humanos, la dignidad humana y la sostenibilidad medioambiental. Además, se enfatiza la importancia de principios fundamentales como la rendición de cuentas y el Estado de derecho. Asimismo, esta declaración abarca capítulos específicos que abogan por una gobernanza más efectiva de los datos, así como por la inclusión y la igualdad de género.

La aplicación de la ética a la IA debe estar presente para cualquier persona, sea de naturaleza pública o privada, que participe en cualquier etapa del ciclo de vida de un sistema de IA, incluyendo la investigación, concepción y el desarrollo y su utilización, además del mantenimiento, el funcionamiento, la comercialización, la financiación, el seguimiento y la evaluación, la validación, el fin de la utilización, el desmontaje y la terminación[29].

---

[27] Fue respaldada por unanimidad por los 193 Estados Miembros de la UNESCO en noviembre de 2021.
[28] Vid. Informe UNESCO sobre ética de la IA. https://www.unesco.org/es/articles/adopcion-del-primer-instrumento-normativo-mundial-sobre-la-etica-de-la-inteligencia-artificial
[29] Resolución del Parlamento Europeo, de 20 de octubre de 2020, sobre un marco de los aspectos éticos de la inteligencia artificial, la robótica y las tecnologías conexas: 2020/2012(INL). Comisión Europea, ver Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones " *Generar confianza en la inteligencia artificial centrada en el ser humano"*

Referido a la cuestión de género, el informe de la UNESCO sobre «Inteligencia Artificial e Igualdad de Género», ofrece un marco para una IA que favorezca la igualdad de género y el bienestar de las mujeres[30]. Plantea la necesidad de reforzar una serie de aspectos. Entre ellos, aumentar la participación y el liderazgo de las mujeres en la educación, la investigación, el desarrollo y la aplicación de la IA, para que sus voces, experiencias y necesidades se tengan en cuenta.

Respecto de esta cuestión, el problema tiene su base en la falta de representación de las mujeres en los campos de ciencia, tecnología, ingeniería y matemáticas (STEM). Según un informe de la UNESCO, actualmente solo uno de cada tres investigadores es mujer, y la representación femenina en la educación superior en carreras STEM se sitúa en poco más del 35%. Además, las mujeres representan menos del 30% de los investigadores científicos a nivel mundial. Este desequilibrio no solo afecta la diversidad y la equidad en estos campos, sino que también tiene implicaciones en la innovación y el desarrollo económico[31]. El dato de que el 78 % de los profesionales de IA son hombres[32], hace que las experiencias masculinas configuren y dominen la creación de algoritmos. Este sesgo de género puede tener consecuencias adversas importantes para las mujeres.

Para abordar esta disparidad, se han implementado diversas iniciativas a nivel nacional e internacional dirigidas a formar, atraer y promover la participación de las mujeres y las niñas en STEM. Estas incluyen programas de mentoría, becas, campañas de sensibilización y esfuerzos para cambiar la percepción cultural sobre las mujeres en estos campos. Es crucial continuar con estos esfuerzos y crear entornos inclusivos que permitan a las mujeres prosperar en las disciplinas STEM, garantizando así que sus talentos y perspectivas contribuyan plenamente al avance de la ciencia y la tecnología.

Otro aspecto señalado es la necesidad de garantizar que los datos, los algoritmos y los dispositivos de la IA sean transparentes, auditables y responsables, y que no contengan ni reproduzcan sesgos de género. Los documentos de la UNESCO, incluido el informe de 2019 «*I'd Blush if I Could: closing gender divides in digital skills through educatio*n», muestran sin ambigüedad que los sesgos de género que persisten en los conjuntos de datos, algoritmos y dispositivos de capacitación de la IA tienen el potencial de propagar y reforzar estereotipos de género perjudiciales.

También es preciso hacer referencia a los acuerdos internacionales que se han tomado recientemente sobre el marco regulador general de la IA. En concreto la «Declaración de Bletchely Park» y el «Proceso de Hiroshima».

La Declaración de Bletchley Park es un hito histórico en la cooperación internacional sobre la inteligencia artificia, que reconoce los beneficios y los riesgos

---

en: *Generar confianza en la inteligencia artificial centrada en el ser humano* [COM(2019) 168].; Grupo de expertos de alto nivel sobre inteligencia artificial, *Directrices éticas para una IA fiable*, 2019.; Grupo de expertos de alto nivel sobre inteligencia artificial, *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment* (*Lista de evaluación para una inteligencia artificial fiable con fines de autoevaluación*), 2020.; La Alianza de la IA es un foro de múltiples interesados que se creó en junio de 2018. Para más información, véase el siguiente enlace: https://ec.europa.eu/digital-single-market/en/european-ai-alliance.; Grupo de expertos de alto nivel sobre IA, *Directrices éticas para una IA fiable*, 2019.

[30]Vid "*Nuevo informe de la UNESCO sobre IA e igualdad de género".* https://www.unesco.org/es/articles/nuevo-informe-de-la-unesco-sobre-inteligencia-artificial-e-igualdad-de-genero

[31] Ver: *"Educación de niñas y mujeres en ciencia, tecnología, ingeniería y matemáticas (STEM)".* En: https://www.unesco.org/es/gender-equality/education/stem

[32]Vid datos del *Global Gender Gap Report* de 2023. https://www3.weforum.org/docs/WEF_GGGR_2023.pdf

que plantea esta tecnología para la humanidad y el planeta. La declaración fue firmada por los representantes de 28 países, entre ellos la Unión Europea, Estados Unidos y China, que son los principales actores en el desarrollo y el uso de la IA. La declaración se basa en los principios y valores compartidos de la democracia, el Estado de Derecho, los derechos humanos y la sostenibilidad, y se compromete a trabajar conjuntamente para garantizar una IA segura, centrada en el ser humano, confiable y responsable.

La *Declaración de Bletchley Park* es el primer paso de un proceso continuo de cooperación global sobre la IA, que tendrá continuidad con futuras cumbres cada seis meses en diferentes países. La próxima cumbre tendrá lugar en Corea del Sur, y después en Francia. El propósito de estas cumbres es crear un panel al estilo del IPCC del cambio climático, que evalúe periódicamente los avances y amenazas de la IA, y que emita recomendaciones y orientaciones para los gobiernos, las organizaciones internacionales, el sector privado, la sociedad civil y la comunidad científica[33].

Por último, el Grupo de los Siete (G7) puso en marcha el pasado mayo de 2023 en la cumbre de Hiroshima (oeste) una iniciativa que ha derivado en la elaboración de once principios para empresas y organizaciones involucradas en el desarrollo y el uso de sistemas de IA. Estos principios se basan en los valores compartidos por el G7, como la democracia, el estado de derecho, el respeto a los derechos humanos y la diversidad.

El código de conducta se denomina el «Proceso de IA de Hiroshima» y que tiene como objetivo promover el desarrollo de sistemas de IA seguros y fiables a nivel internacional y gestionar sus riesgos. El código de conducta se revisará periódicamente para adaptarse a la rápida evolución de la tecnología y se cooperará con los actores implicados en el sector de la IA, como la sociedad civil, la academia, el sector privado y las organizaciones internacionales. El código de conducta fue respaldado por los líderes del G7 (Alemania, Canadá, Estados Unidos, Francia, Italia, Japón y el Reino Unido) el 30 de octubre de 2023.

La pregunta que se plantea es si está presente la perspectiva de género en estos documentos. Pues bien, no se prioriza la cuestión de género, ni se hace referencia específica a ella. En los casos en que sí se hace referencia, no se explica cómo llevar a cabo la aplicación.

La Carta de Derechos Digitales española presentada el 14 de julio de 2021 (De la Sierra Morón, 2022); (Barrio Andrés, 2021); (Barrio Andrés, 2022) si incluye una referencia a la cuestión de género. Se trata de un documento sin carácter normativo y se configura como un marco de referencia para guiar futuros proyectos legislativos, así como el desarrollo de las políticas públicas. Quiere sentar principios que guíen futuros proyectos legislativos y el desarrollo de las políticas públicas. Su objetivo es reforzar los derechos de la ciudadanía, generar certidumbre en la nueva realidad digital y aumentar la confianza ante los cambios y disrupciones tecnológicas (Cotino Hueso, 2022).

La Carta de Derechos Digitales reconoce el derecho a la igualdad y no discriminación. Se reconocen el derecho y el principio a la igualdad, como inherentes a las personas y aplicables en los entornos digitales, incluyendo la no discriminación y la no exclusión. En particular, se promoverá la igualdad efectiva de mujeres y

[33] Vid. "*Documento normativo: Declaración de Bletchley de los países que asisten a la Cumbre de seguridad de la IA, 1 y 2 de noviembre de 2023".* https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/dbc58681-1b68-47e0-8e3f-f91435fdf8ce

hombres en entornos digitales. Se fomentará que los procesos de transformación digital apliquen la perspectiva de género[34].

La autorregulación se erige, también, como un pilar fundamental en la aplicación de los principios de inteligencia artificial, se refiere a la capacidad de las empresas y organizaciones para establecer sus propias normas y directrices éticas en el desarrollo y uso de tecnologías de IA. Este enfoque se basa en la idea de que las entidades que crean y utilizan IA son las más adecuadas para entender sus implicaciones y, por lo tanto, para regular su comportamiento de manera responsable. Sin embargo, la autorregulación no está exenta de críticas. Algunos expertos argumentan que la autorregulación por sí sola puede no ser suficiente para abordar todos los desafíos éticos y sociales que presenta la IA. La preocupación principal es que sin una supervisión externa, las empresas podrían priorizar sus intereses comerciales sobre las consideraciones éticas, lo que podría llevar a abusos o a la implementación de sistemas de IA con sesgos o efectos negativos no intencionados[35].

El intento de autorregulación que ha tenido un mayor impacto es el de la «Conferencia de Asilomar sobre IA beneficiosa». Tuvo lugar en California, organizada por *Future of Life Institute* en enero de 2017, se formulan principios para una IA ética. Los veintitrés principios están divididos en temas o preguntas de investigación; temas concernientes a la ética y valores y, problemas a largo plazo.

Dentro de los principios de valores humanos menciona específicamente que los sistemas de IA deben ser diseñados y operados de manera que sean compatibles con los ideales de dignidad humana, derechos, libertades y diversidad cultural. Esto implica un reconocimiento de la necesidad de considerar y respetar la diversidad de género en el diseño y la implementación de sistemas de IA[36].

Por otra parte, Google ha establecido siete principios éticos para guiar el uso de la IA, entre los cuales se incluye el compromiso de evitar la creación o el refuerzo de sesgos injustos. Esto implica trabajar activamente para prevenir impactos adversos en las mujeres[37]. Adicional a estos siete principios, Google garantiza que no diseñará o desplegará desarrollos en inteligencia artificial en ciertas áreas, como aquellas que puedan causar daños o vulnerar normas o principios de derecho internacional ampliamente aceptados y derechos humanos.

De manera similar, Microsoft enfatiza la importancia de la equidad y la inclusión en sus sistemas de inteligencia artificial, asegurando que todos los

---

[34] Vid. Real Decreto 22/2023, por el que se aprueba el estatuto de la Agencia Española de Supervisión de la Inteligencia Artificial (AESIA). La AESIA se adscribe al Ministerio de Asuntos Económicos y Transformación Digital a través de la Secretaría de Estado de Digitalización e Inteligencia Artificial. Con la creación de esta Agencia, España se convierte en el primer país europeo en tener un órgano de estas características y se anticipa a la entrada en vigor del Reglamento Europeo de Inteligencia Artificial. Dicho proyecto de reglamento establece para los Estados miembros la obligación de seleccionar una autoridad nacional de supervisión que se encargue de supervisar la aplicación de la normativa en materia de Inteligencia Artificial; Vid. Real Decreto 729/2023, de 22 de agosto, por el que se aprueba el Estatuto de la Agencia Española de Supervisión de Inteligencia Artificial: https://www.boe.es/diario_boe/txt.php?id=BOE-A-2023-18911; Texto del estatuto: https://www.hacienda.gob.es/Documentacion/Publico/GabineteMinistro/Notas%20Prensa/2023/CONSEJO-DE-MINISTROS/22-08-23-NP-CM-Estatutos-Agencia-Inteligencia-Artificial.pdf

[35] Vid informe "Inteligencia artificial: un equilibrio entre la regulación y la autorregulación Texto original de Ryan Hagemann y Jean-Marc Leclerc, con adaptaciones del equipo de Relaciones Gubernamentales de IBM América Latina". https://ia-latam.com/wp-content/uploads/2020/01/IBM-Inteligencia-artificial-un-equilibrio-entre-la-regulaci%C3%B3n-y-la-autorregulaci%C3%B3n.pdf

[36] Vid la Web "*futureoflifle*" en: https://futureoflife.org/

[37] Vid informe sobre estadísticas de Google de 2022. https://ai.google/static/documents/ai-principles-2022-progress-update.pdf

individuos sean tratados de manera justa y equitativa[38]. Microsoft destaca una serie de principios rectores para la inteligencia artificial responsable: responsabilidad, inclusión, confiabilidad y seguridad, equidad, transparencia y privacidad y seguridad. Además, actuará de forma que mitigue los sesgos sociales. Asegurándose de que el idioma y el comportamiento no presentan estereotipos ni sesgos no deseados. Plantea, como ejemplo, una característica de autocompletar debe ser inclusiva de la identidad de género.

IBM incluye entre sus obligaciones la necesidad de realizar pruebas de sesgos. Parte de que las organizaciones involucradas en el desarrollo del ciclo de vida de la IA comparten cierto nivel de responsabilidad en asegurar que los sistemas que diseñan sean imparciales y seguros. Y como con cualquier software para uso comercial, es esencial realizar pruebas continuas de distintos tipos (protección de datos, cumplimiento, antidiscriminación, protección del consumidor, seguridad, etc.), para identificar y reducir las posibilidades de que el aprendizaje automático produzca resultados no deseados[39].

La regulación formal puede establecer estándares mínimos y garantizar que se respeten los derechos fundamentales, mientras que la autorregulación puede permitir una mayor flexibilidad y adaptabilidad a las rápidas innovaciones tecnológicas.

En resumen, la autorregulación es un componente importante en la aplicación de los principios de IA, pero debe complementarse con regulaciones claras y precisas para asegurar que la IA se desarrolle y utilice de manera ética y beneficiosa para la sociedad.

5. Conclusiones

Los estudios recientes han puesto de manifiesto un sesgo de género inherente en los sistemas de inteligencia artificial (IA), que perpetúa y amplifica las desigualdades existentes. Para contrarrestar esta tendencia, es imperativo implementar estrategias de construcción de género que sean holísticas y transversales, abarcando todos los estratos sociales y promoviendo la igualdad de género de manera activa y tangible.

Para lograr un empoderamiento efectivo de las mujeres en este contexto y evitar la reproducción y el refuerzo de estereotipos y acciones discriminatorias es fundamental utilizar estrategias integrales de construcción de género. Esto implica adoptar un enfoque de transversalidad que abarque todos los niveles de la sociedad y que promueva la igualdad de género no solo en teoría, sino en la práctica, a través de acciones concretas y positivas.

Es esencial que los sistemas de IA sean sometidos a una intervención consciente y deliberada en todas las fases de su desarrollo, desde la concepción hasta la implementación. La adopción de principios regulatorios como la transparencia y la responsabilidad es crucial para garantizar que la IA avance hacia la equidad de género.

[38] Vid la referencia al diseño inclusivo de Microsoft en: https://inclusive.microsoft.design/
[39] Vid informe "Inteligencia artificial: un equilibrio entre la regulación y la autorregulación Texto original de Ryan Hagemann y Jean-Marc Leclerc, con adaptaciones del equipo de Relaciones Gubernamentales de IBM América Latina": https://ia-latam.com/wp-content/uploads/2020/01/IBM-Inteligencia-artificial-un-equilibrio-entre-la-regulaci%C3%B3n-y-la-autorregulaci%C3%B3n.pdf

Resulta preocupante que, ante la irrupción de una transformación tan radical como la que supone el desarrollo de sistemas IA, toda la atención se ha centrado en estudiar sus efectos. De tal forma que la perspectiva de género parece que haya sido relegada a un segundo plano en el estudio de sus implicaciones. Es vital que el análisis de género se integre de manera transversal en todas las áreas afectadas por la IA, dada su influencia omnipresente en la vida cotidiana.

Para mitigar el sesgo de género, se deben adoptar prácticas que incluyan una selección de datos diversa y representativa, así como una revisión constante de los algoritmos para identificar y corregir prejuicios. Estas medidas son fundamentales para fomentar una IA equitativa e inclusiva, que sea aplicable desde una perspectiva amplia e interseccional de la sociedad.

Bibliografía

Barrio Andrés, M. (2021). Génesis y desarrollo de los derechos digitales. *Revista de Las Cortes Generales,* 110, 197-233. https://doi.org/10.33426/rcg/2021/110/1572

Barrio Andrés, M. (2022). Garantía de los derechos en los entornos digitales. En Cotino Hueso, L. (Coord.), *La Carta de Derechos Digitales* (363-395). Tirant Lo Blanch.

Boix Palop, A. (2020) Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones. *Revista de Derecho Público: teoría y método,* 1, 223-269.

Buolamwini, J. y Timnit, G. (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research Conference on Fairness, Accountability, and Transparency,* 1-15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Coltrane, S. y Messineo, M. (2000). The Perpetuation of Subtle Prejudice: Race and Gender Imagy in 1990s Television Advertising. *Sex Role,* (42), 363-389.

Cotino Hueso, l y Castellanos Claramunt, J. (2022). Transparencia y Explicabilidad de la Inteligencia Artificial y "Compañía" (Comunicación, Interpretabilidad, Inteligibilidad, Auditabilidad, Testabilidad, Comprobabilidad, Simulabilidad...). Para qué, para quién y cuánta. *Transparencia y explicabilidad de la Inteligencia Artificial.* Tirant Lo Blanch.

Cotino Hueso, L. (2022). *La Carta de Derechos Digitales*, Tirant Lo Blanch.

De la Sierra Morón, S. (2022). Una introducción a la carta de derechos digitales. En Cotino Hueso, L. (Coord.), *La Carta de Derechos Digitale*s (27-52). Tirant Lo Blanch.

Gutiérrez David, M. (2021). Administraciones inteligentes y acceso al código fuente y los algoritmos públicos. Conjurando riesgos de cajas negras decisionales. *Derecom*, 31, 19-105.

Hernández Peña, J. C. (2022). *El marco jurídico de la inteligencia artificial. Principios, procedimientos y estructuras de gobernanzas.* Aranzadi.

Hundt, A., Agnew, W., Zeng. W., Kacianka, S. y Gombolay, M. (2022*).* Robots Enact Malignant Stereotypes*. ACM Conference on Fairness, Accountability, and Transparency FAccT '22, June 21–24.* https://doi.org/10.1145/3531146.3533138.

Kim, J. L., Sorsoli, C. L., Collins, K., Zylbergold, B. A., Schooler, D., y Tolman, D. L. (2007). From Sex to Sexuality: Exposing the Heterosexual Script on Primetime Network Television. *The Journal of Sex Research*, 44 (2), 145–157. http://www.jstor.org/stable/25701753.

Kullmann, M. (2018). *Platform* Work, Algorithmic Decision-Making, and EU Gender Equality Law. *International Journal of Comparative Labour Law and Industrial Relations*, 34 (1), 1-21.

Lovdal, L. (1989) Sex Role messages in television commercials: An updat. *Sex Roles*, 21, 715-724. https://doi.org/10.1007/BF00289804.

Martín Delgado, I. (2009). La gestión electrónica del procedimiento administrativo. *QDL*, 21, 84-101.

Martín Delgado, I. (2009). Naturaleza, concepto y régimen jurídico de la actuación administrativa automatizada. *Revista de Administración Pública*, 180, 353-386.

Matthes, J., Prieler, M. y Adam, K. (2016). Gender-role portrayals in television advertising across the globe. *Sex Roles. A Journal of Research,* 75 (7-8), 314-327. https://doi.org/10.1007/s11199-016-0617-y.

Mehrabi, N., Morstatter, F., Saxena, N. Lerman, K. y Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *arXiv* 2022. https://doi.org/10.48550/arXiv.1908.09635

Serra Cristobal, R. (2004) La discriminación indirecta por razón de sexo, en Ridaura Martínez, M.J. y Aznar Gómez, M. (Coords.), Discriminación versus Diferenciación (Especial referencia a la problemática de la mujer) (365-398). Tirant lo Blanch.

Vestri, G. (2021). La inteligencia artificial ante el desafío de la transparencia algorítmica: Una aproximación desde la perspectiva jurídico-administrativa. *Revista Aragonesa de Administración Pública*, 56, 368-398.