Artigo Original

Encontros Bibli

LOS ESTEREOTIPOS DE CLASE, GÉNERO Y RAZA REPRODUCIDOS POR LA IA GENERATIVA: RECOMENDACIONES PARA LOS USUARIOS

Class, gender and race stereotypes reproduced by generative AI: recommendations for users

Andrea Castro-Martínez
Universidad de Málaga
Dpto. Comunicación Audiovisual y Publicidad
Málaga, España
andreacastro@uma.es
https://orcid.org/0000-0002-2775-625X

José-Luis Torres-Martín
Universidad de Málaga
Dpto. Comunicación Audiovisual y Publicidad
Málaga, España
jltorres@uma.es
https://orcid.org/0000-0001-6556-1560

Cristina Pérez-Ordóñez
Universidad de Málaga
Dpto. Comunicación Audiovisual y Publicidad
Málaga, Españacristinaperezordonez@uma.es
https://orcid.org/0000-0002-9532-0087

A lista completa com informações dos autores está no final do artigo

RESUMEN

Objetivo: Establecer recomendaciones para que los prosumidores efectúen un uso ético de estas herramientas evitando, de esta forma, resultados que caigan en estereotipos de clase, raza y/o género.
Método: Análisis del material documental desarrollado por distintas entidades: instituciones internacionales (UNESCO y Conuncil of Europe) y empresas desarrolladoras de herramientas de IA (OpenAI, Google, Microsoft y Anthropic).
Resultado: La IA puede perpetuar estereotipos ya que emplea datos que reproducen un contexto donde las minorías y las mujeres están infrarrepresentadas. Las principales corporaciones afirman haber desarrollado códigos de conducta y autorregulación. Esta gobernanza concuerda con las recomendaciones internacionales y rige su comportamiento, los protege de responsabilidades y basa su RSC. Incluyen recomendaciones para los usuarios en los procesos de uso de las herramientas.
Conclusiones: Aún no es posible evitar totalmente los sesgos, pero es recomendable que los usuarios detecten los estereotipos y los minimicen. La autorregulación contribuye a contenidos menos discriminatorios, además de incorporar a mujeres y minorías como desarrolladores. Plantean códigos de conductas para detectar contenidos inapropiados. Los hallazgos son útiles para fomentar la perspectiva crítica y poner el foco en el papel que los usuarios, a los que se ofrecen recomendaciones, tienen en un proceso en el que el uso que se haga de las herramientas resulta fundamental.
PALABRAS CLAVE: Inteligencia artificial. Género. Raza. Clase. Estereotipos.

ABSTRACT

Objective: To establish recommendations for prosumers to make ethical use of these tools, thus avoiding results that fall into class, race and/or gender stereotypes.
Method: Analysis of documentary material developed by different entities: international institutions (UNESCO and Council of Europe) and companies developing AI tools (OpenAI, Google, Microsoft and Anthropic).
Finding: AI can perpetuate stereotypes as it uses data that reproduces a context where minorities and women are underrepresented. Major corporations claim to have developed codes of conduct and self-regulation. This governance is in line with international recommendations and governs their behaviour, protects them from liability and underpins their CSR. They include recommendations for users in the processes of using the tools.
Conclusions: It is not yet possible to completely avoid biases but it is advisable for users to detect stereotypes and minimise them. Self-regulation contributes to less discriminatory content, as well as incorporating women and minorities as developers. They propose codes of conduct to detect inappropriate content. The findings are useful to encourage a

critical perspective and focus on the role of users, to whom recommendations are offered, in a process in which the use of tools is fundamental.
KEYWORDS: Artificial intelligence. Gender. Race. Class. Stereotypes.

1 INTRODUCCIÓN

La ecología de los medios y de las relaciones humanas se ha visto alterada en las últimas décadas por los procesos de digitalización (Arjona-Martín; Méndiz-Noguero; Victoria-Más, 2020). En este nuevo estado de las cosas, más complejo y fluctuante (KIM, KANG, & LEE, 2021), las prácticas que tienen lugar en la sociedad están determinadas en gran medida por la cibercultura (LEVY, 2010; DURMUS, 2021; TURNER, 2021).

De esta forma, la posición de los usuarios, más pasiva en el paradigma anterior, se ha transformado hasta convertirse en prosumidores, esto es, productores y consumidores de productos culturales (Kotler, 2010; Jenkins; Ford; Green, 2013). Pese a ello, e igual que ocurría en el marco comunicativo anterior (Curran, 2005), esta cultura digital reproduce la ideología de la clase dominante (Van Dijk, 2011; Fuchs, 2021; Piñeiro-Otero; Martínez-Rolán, 2021).

Esto contradice la hipotética democratización que iba a suponer la era digital con eventos tales como la construcción colectiva o la visibilización de esferas públicas alternativas (Joyce, 2010; Carty, 2015; Sosa; Galarza; Castro-Martinez, 2019).

Diferentes instituciones y organizaciones internacionales, como es el caso del Council of Europe (2023), la UNESCO (2023) o Equality Now (2023), han puesto de manifiesto que los algoritmos empleados por la IA pueden generar desigualdades entre hombres y mujeres, entre clases sociales y económicas y entre individuos racializados (Cave; Dihal, 2020; Lamensch, 2023). Esto se debe a que reafirman prácticas discriminatorias al emplear datos procedentes de la red, que pueden no ser ciertos (Parra-Valero; Oliveira, 2018) o en los que las representaciones de estos colectivos no se encuentren equilibradas (Roekens, 2023).

Por otra parte, cualquier fase de desarrollo de los sistemas de IA está sujeta a reproducir desigualdades (Manasi; Panchanadeswaran; Sours, 2023) si no se pone el foco en minimizar los efectos de las voces dominantes y en incorporar perspectivas que ofrezcan una diversidad real.

La presente investigación adopta una posición abiertamente crítica desde la perspectiva de clase, feminista y de raza para identificar las desviaciones y estereotipos que pueden reproducirse a través del uso de la IA, con el objetivo de promover puntos de

vista más críticos, inclusivos y equitativos por parte de los usuarios. Este trabajo parte de dos preguntas de investigación que guían tanto la elaboración del marco teórico, que contextualiza la importancia y uso de la IA y la influencia de los estereotipos, como del análisis de diferentes actores implicados en su desarrollo y supervisión. Del mismo modo, sirven de eje para establecer sus objetivos propositivos y pedagógicos. Estas preguntas son las siguientes:

- Pregunta de investigación 1- ¿Se puede evitar la reproducción de estereotipos de clase, género y raza en los productos generados a través de la IA? (PI1)

- Pregunta de investigación 2- De ser así, ¿cómo pueden los usuarios detectarlos y minimizar su aparición? (PI2).

Para dar respuesta a estas dos cuestiones se diseña una investigación exploratoria con distintos objetivos:

- El Objetivo General de la misma (OG) es establecer recomendaciones para que los prosumidores efectúen un uso ético de estas herramientas evitando, de esta forma, resultados que caigan en estereotipos de clase, raza y/o género. Para alcanzar esta meta se establecen también distintos objetivos secundarios que se concretan del siguiente modo:

- Como Primer Objetivo Secundario (OS1) se pretende revisar la documentación generada por las cuatro grandes empresas del sector -Microsoft, OpenAI, Google y Anthropic-, que se han propuesto autorregular sus modelos de IA (Europa Press PortalTIC, 2023).

- Asimismo, el Segundo Objetivo Secundario busca detallar los códigos de buenas prácticas que declaran las instituciones internacionales al respecto (OS2).

- Por último, el Tercer Objetivo Secundario pretende describir los problemas y procesos que se plantean en cuanto a la interacción de los propios usuarios con la IA (OS3).

Para alcanzar estas metas se ofrece un análisis de los sistemas de producción de contenidos mediante estas tecnologías, así como una serie de recomendaciones para mejorar los resultados obtenidos.

2 DISCRIMINACIÓN, INTERSECCIONALIDAD Y REPRESENTACIONES MEDIÁTICAS

Las representaciones mediáticas de las mujeres y de las minorías han sido ampliamente estudiadas por la influencia que ejercen sobre los estereotipos. El motivo es

que los medios de comunicación son un agente socializador de gran poder (Brooks; Hébert, 2006) que contribuye en buena medida a conformar la imagen que la ciudadanía tiene de las sociedades en las que habita, así como de las normas que la rigen.

Los estereotipos son representaciones que asignan a determinados colectivos características, roles y modelos de conducta a los que se deben sus integrantes, como ocurre con los asignados a hombres y mujeres (Lagarde, 1998; Freixas, 2000). Los imaginarios sociales condicionan las matrices conceptuales y perceptuales que se emplean para entender e interpretar la realidad (Castoriadis, 1983), por lo que los estereotipos se emplean en los procesos de socialización y comunicación entre los grupos.

En el caso de mujeres que pertenecen a colectivos que sufren múltiples discriminaciones, estos estereotipos se acentúan (Bruno, 2016; Colella; Gianturco, 2020). El racismo de género tiene influencia en cuestiones tan dispares como el acceso a la sanidad y la salud sexual y reproductiva por parte de las mujeres negras y latinas (Rosenthal; Lobel, 2018) o el acceso limitado a recursos económicos y a oportunidades de desarrollo, por lo que estas mujeres se enfrentan a desafíos únicos debido a su subordinación social (Spates et al., 2020). Y es que la desigualdad se incrementa en el caso de la interseccionalidad y la discriminación múltiple:

> El concepto de interseccionalidad viene referido a la discriminación múltiple que sufren las mujeres, bien por el hecho de ser mujeres sumándole los motivos de raza, etnia, orientación sexual, discapacidad, etc. La convergencia de los estereotipos aumenta la probabilidad de sufrir una situación de discriminación (Sánchez; García; Japa, 2022, p.71).

La activista por los derechos civiles Angela Davis resulta una autora de referencia al incorporar a la discriminación racial el feminismo, la defensa de los derechos LGTBIQ+ y el abolicionismo del sistema carcelario, entre otras cuestiones, desde posicionamientos marxistas (Davis,1981, 2011). Tanto en su autobiografía (Davis, 2017) como en su abundante producción académica aborda la desigualdad racial y de género patente en la sociedad estadounidense desde una perspectiva interseccional.

La construcción indentitaria de las audiencias de cualquier mensaje y, en especial de los espectadores de los medios de comunicación, tanto a nivel individual como colectivo, se ve profundamente influenciada por las informaciones que consume (Cabrera, 2004). La imagen es la gran protagonista de la época actual y con ella se identifican las audiencias

(Sola-Morales, 2014) cuando reciben impactos de los distintos soportes, de modo que los imaginarios sociales (Castoriadis, 1983) condicionan la forma en que se entienden, interpretan y reinterpretan.

Estas imágenes contribuyen a la generación de estereotipos dominantes en un contexto determinado en el que se utilizan como vía de socialización para sus miembros, pues para pertenecer al grupo concreto es común que deban aceptarlos e identificarse con ellos (González, 1999).

Esto ocurre, por ejemplo, en el caso de las mujeres, puesto que, en palabras de Castejón:

> la imagen y su representación ha sido uno de los mecanismos culturales más poderosos y efectivos de control social sobre las mujeres, ya que las representaciones transmiten pautas de comportamiento y roles sociales que las mujeres interiorizan y asumen (Castejón, 2004, p. 311).

En esta interiorización, asimilación y replicación de roles y comportamientos resultan claves los medios de comunicación y los productos culturales que consume una sociedad. El modo en que los medios de comunicación construyen la imagen de los colectivos ha sido y continúa siendo un eje de las luchas populares y académicas, tanto desde la perspectiva de género (Byerly; Ross, 2008) como de las personas racializadas (Scharrer; Ramasubramanian, 2015) y de las luchas de clase (Hesmondhalgh, 2017).

Esto se debe a que lo habitual es que estos colectivos no ostenten el mismo protagonismo en los medios de comunicación que las mayorías, principalmente conformadas por hombres blancos, normalmente occidentales (THORNHAM, 2007; ADCOCK, 2010; DOUKHAN et al., 2018). Es el caso de los estereotipos de género, que han sido ampliamente analizados (Lagarde, 1998; Freixas, 2000) y cuya función es fijar atributos y roles con la función de que hombres y mujeres se identifiquen con ellos y su conducta se rija por los modelos que consumen. Lo mismo ocurre con las personas racializadas, ya que los mensajes aportados por los medios inciden en el desarrollo de las cogniciones raciales y étnicas y en los comportamientos intergrupales (Mastro, 2009; Tukachinsky; Mastro; Yarchi, 2017). Así, por ejemplo:

> la categorización basada en la clase social mediatiza el efecto de la desigualdad y los estereotipos de clase social en la dimensión vertical -asertividad y competencia-, y prueba que las relaciones intergrupales negativas (es decir, el aumento de la

> competencia y la reducción de la confianza) explican los estereotipos en la dimensión horizontal -amabilidad y moralidad-" (Tanjitpiyanond; Jetten; Peters, 2022).

Otro aspecto relevante en este sentido es que la representación de las mujeres, las clases bajas y las minorías es menos ajustada a la realidad que la de otros sujetos y actores con mayor poder (Bligh et al., 2012; Cheryan et al., 2013; Ramasubramanian; Banjo, 2024). Dukes; Gaither (2017), por su parte -en un estudio sobre la representación mediática de las minorías raciales y étnicas de las víctimas de la violencia policial y su impacto en las sentencias penales- ponen como ejemplo que el tipo de información difundida sobre una víctima puede influir significativamente en las actitudes hacia la misma y hacia el tirador.

Cada vez más las mujeres y otras minorías ostentan mayor poder y mayor número de puestos creativos y de decisión en la creación de productos culturales, lo que afecta tanto a las temáticas abordadas como al modo de tratarlas (Ruffinelli, 2014; Monedero; Impelluso, 2021; Kendall, 2011; Torres-Martín et al., 2022; Cuenca Orellana, 2022).

No obstante, en la industria cultural y mediática prevalece la perspectiva masculina (Ruiz-Muñoz; Sánchez, 2008; Torres-Martín; Castro-Martínez; Pérez-Ordóñez, 2023), por lo que se hace necesario realizar un planteamiento que reintegre el análisis de clase en los modelos sociales de raza, sexualidad y género (Munt, 2000).

Debido a la relevancia de la Inteligencia Artificial ha adquirido en la actualidad y a las previsiones de su gran impacto futuro, resulta imprescindible tener consciencia de que en su desarrollo están presentes los estereotipos de clase, raza y género que se encuentran de forma mayoritaria en la red y en gran parte de las sociedades contemporáneas (West; Whittaker; Crawford, 2019). Su veloz desarrollo no debe opacar la necesidad de estudiar el modo de integrar las cuestiones éticas, el bienestar individual y grupal, la agencia moral y la prevención de daños, además de una toma de decisiones democrática (Dubber; Pasquale; Das, 2020).

En este sentido es necesario tener en cuenta tanto el diseño de las herramientas como el uso de las mismas por parte de quien las emplee (Ruggiero, 2000). Autores como McQuillan (2022) o Eksi plantean los peligros de la IA para la democracia, la ciudadanía y alertan de su uso por parte de posicionamientos de extrema derecha en el entorno digital. También se ha señalado su papel en el criptofascismo (Pinto, 2019; Mussa, 2019) y en la manosfera, donde se producen contenidos falsos como manifestaciones de la violencia

política de género (Marwick; Caplan, 2018; Horta et al., 2021; Bosch; Guillem; López, 2022; Barrientos-Báez; Piñeiro; Porto, 2024).

Un antifeminismo altamente tóxico ha cobrado protagonismo en una serie de redes y plataformas en línea (Ging; 2019; Bazzano, 2023) donde el uso de los *deepfakes* y de los contenidos generados por Inteligencia Artificial han proliferado y se ha extendido hasta la esfera física, donde se han empezado a difundir imágenes de desnudos de mujeres generados con esta tecnología (Contreras, 2024).

3 PROCEDIMIENTOS METODOLÓGICOS

La metodología que se establece para alcanzar los objetivos mencionados se centra en el análisis del material documental desarrollado por distintas entidades: instituciones internacionales y empresas desarrolladoras de herramientas de IA.

Distintas organizaciones e instituciones a nivel internacional están trabajando para detectar y denunciar posibles sesgos producidos por la implementación de la IA en cualquier fase de su elaboración. Algunos casos sobre los que se han estudiado sus efectos perversos son, por ejemplo, la hipersexualización de las imágenes femeninas, la discriminación en procesos selectivos aplicados en los servicios de recursos humanos de las empresas (Lamensch, 2023) o la discriminación de distintos colectivos en cuanto a la representación de determinados trabajos (Manasi; Panchanadeswaran; Sours, 2023).

Por su parte, las empresas desarrolladoras de IA han incorporado dentro de sus planteamientos de Responsabilidad Social Corporativa (RSC) los principios éticos de igualdad y equidad aplicados a todos sus procesos productivos. Por este motivo las grandes corporaciones de esta industria se han propuesto autorregular sus modelos generativos para evitar desviaciones que puedan derivar en efectos negativos sobre la sociedad.

La Academia no es ajena a esta problemática y en los últimos años se ha incrementado el número de estudios que profundizan sobre la reproducción de estereotipos generados por estas nuevas tecnologías (Adams; Loideáin, 2019; Craiut; Iancu,2022) y por el empleo que hacen los prosumidores de las mismas (Fox, 2018; Ahn; Kim; Sung, 2022).

Las entidades, organizaciones y personas que conforman la muestra del estudio son las siguientes:

- Instituciones internacionales: Council of Europe y UNESCO. Se analizan los estudios publicados por estas entidades sobre ética aplicada a la IA.

- Principales empresas desarrolladoras de IA: OpenAI, Google, Microsoft y Anthropic. Se analizan sus distintas declaraciones éticas.

Se ha optado por la revisión bibliográfica documental (Ahmed, 2010), ya que permitirá conocer desde un prisma múltiple las distintas aristas de este objeto de estudio. En cuanto a la tipología de estereotipos tenidos en cuenta son aquellos que las entidades reconocen derivados de la clase, el género y la raza, teniendo en cuenta su posible interseccionalidad.

4 RECOMENDACIONES INTERNACIONALES SOBRE UNA IA MÁS IGUALITARIA

El Consejo de Europa (CE) plantea que la educación debe ser accesible, inclusiva y equitativa y entiende que la IA tiene potencial transformador para alcanzar estos objetivos. Pretende garantizar que la adopción de sistemas de IA esté alineada con los derechos humanos y la dignidad del alumnado y debe hacerse en todo el ciclo de vida de los sistemas, desde el diseño y la modelización hasta la formación.

El CE pone el foco en la importancia de la transparencia, la responsabilidad y la protección de los derechos fundamentales, además de promover un uso responsable y ético fomentando un enfoque inclusivo que respete la diversidad cultural, lingüística y social. Se centra en combatir los prejuicios derivados de la discriminación algorítmica, en lo que cobra especial importancia la diversidad de los equipos. Cualquier sistema creado por un grupo mayoritario corre el riesgo de no integrar las perspectivas de los colectivos marginados, lo que se traduce en algoritmos que sólo funcionan para la mayoría. Por ello recomienda incorporar diversidad que aporte enfoques holísticos que hagan la IA más completa y responsable.

Otra de sus medidas ha sido crear un Comité de Expertos en Inteligencia Artificial, Igualdad y Discriminación (GEC/ADI-AI) que está compuesto por ocho representantes de los Estados miembros y seis expertos independientes. Se trata de un subcomité conjunto de la Comisión de Igualdad de Género (GEC) y del Comité Directivo sobre Antidiscriminación, Diversidad e Inclusión (CDADI); su labor consistirá en redactar para 2025 una Recomendación sobre el impacto de los sistemas de IA, su potencial para promover la igualdad -incluida la igualdad de género- y los riesgos que pueden causar en relación con la no discriminación.

La UNESCO, por su parte, reconoce el peligro que suponen los sesgos que pueden incorporar y exacerbar la IA y que considera una amenaza. Por ello señala que los actores del proceso deben promover la justicia social, la equidad y la lucha contra la discriminación, así como adoptar un enfoque inclusivo que garantice la disponibilidad de los beneficios de esta tecnología.

Para lograrlo pide compromiso por parte de los implicados en su desarrollo para asegurar la auditabilidad, trazabilidad y explicabilidad de los algoritmos, los datos y los procesos de concepción, además de un examen externo de los sistemas para asegurar la transparencia. También plantea la necesidad de que las entidades asuman los principios de responsabilidad y rendición de cuentas, desde perspectivas multidisciplinares, multiculturales, pluralistas e inclusivas.

Cuadro 1. Recomendaciones del Consejo de Europa y de la UNESCO sobre el desarrollo de la IA.

Prevención, transparencia y rendición de cuentas	- Auditabilidad, trazabilidad y explicabilidad de algoritmos, datos y procesos de concepción
Diversidad, inclusión, representación y participación	- Participación democrática, sensibilización pública y desarrollo de capacidades - Promover la justicia social, la equidad y la lucha contra la discriminación y adoptar un enfoque inclusivo
Acceso a la justicia y reparación legal	- Establecer mecanismos para detectar y subsanar errores.

Fonte: Elaboración propia.

5 MEDIDAS ADOPTADAS POR LAS DESARROLLADORAS DE IA

Por su parte, las principales corporaciones implicadas en el desarrollo de la IA han decidido elaborar códigos de conducta y dotarse de una autorregulación que rija su comportamiento y responsabilidades futuras, creando un sistema de gobernanza en base a comités éticos.

OpenAI afirma que para evitar los sesgos ha adoptado distintas medidas: salvaguardar sus contenidos de los prejuicios, prevenir el engaño y conocer la fuente.

Para evitar los prejuicios han incorporado tres medidas:

- Moderación de contenido y filtros de seguridad: entrenan su sistema de IA, implementado filtros que contribuyan a evitar que se generen resultados sesgados y dañinos.

- Prevención de estereotipos en imágenes: disponen de una capa de seguridad que verifica en la creación de imágenes si hay afirmaciones infundadas o cualquier elemento que infrinja la política corporativa en activo.

- Lucha contra estereotipos vocales: entrenamiento del sistema para evitar reforzar prejuicios negativos y diseño que la hace sensible a los acentos, además de abstenerse de hacer suposiciones sobre el origen, la inteligencia o la raza de una persona basándose únicamente en su voz.

Con el objetivo de prevenir el engaño y conocer las fuentes aplican dos principios:

- Los *Deep Fakes*, o recreaciones no autorizadas de imágenes y vídeo generadas mediante IA son inaceptables, ya que suponen una manipulación y tergiversación deliberada de personas o ideas.

- Verificación de autenticidad: han instaurado sistemas que permiten rastrear y verificar la procedencia de las imágenes e identificar su autoría. Para ello se agregan metadatos C2PA a todas las imágenes creadas y editadas por DALL•E 3 en ChatGPT y OpenAI API. Se han asociado con el Comité Directivo de la Coalición para la Procedencia y Autenticidad del Contenido (C2PA) y, aún en fase de prueba, está el clasificador de detección DALL·E 3 para predecir la probabilidad de que una imagen se origine con sus sistemas.

Los procesos que OpenAI aplica para trabajar la seguridad de sus herramientas se basa en ciertos aspectos básicos que luego implementan y mejoran en las diferentes versiones de sus productos. De este modo se incorpora a la sociedad y a los usuarios al desarrollo de esta tecnología, a la que deben adaptarse, de modo que se involucra a las distintas partes interesadas. Se mejora la experiencia y se detectan sesgos en la adopción de esta tecnología:

- Enseñar a la IA sobre el contenido positivo y el dañino, filtrando este y ofreciendo respuestas con empatía.

- Realizar pruebas internas y externas con expertos en el mundo real para mejorar sus sistemas de seguridad y prevención.

- Compartir a través de comentarios reales e incorporar sus sugerencias para hacer sus herramientas más seguras y útiles. Han puesto su API a disposición de los desarrolladores para implementar la tecnología en sus aplicaciones y así poder monitorizar usos indebidos a los que dar respuesta.

- Aprender el uso que se le da a la IA en el mundo real, intentando prevenir riesgos en la fase de implementación, pero también incorporando posteriormente mejoras basadas en los usos y gratificaciones de los prosumidores. Esto facilita la detección de riesgos derivados de determinados comportamientos, así como la incorporación de usos beneficiosos.

Los aspectos éticos con los que se ha comprometido OpenAI son los siguientes:

- Invertir en investigación e ingeniería para minimizar los sesgos y mejorar la equidad y la inclusión en sus modelos.

- Compromiso con la transparencia y la rendición de cuentas, ofreciendo explicaciones de las capacidades y limitaciones de sus herramientas, incorporando auditorías y aportaciones externas para su mejora.

- Compromiso con la lucha contra usos que puedan dañar a la Humanidad o favorezcan las concentraciones indebidas de poder, haciendo valoraciones cuidadosas de los riesgos potenciales y el impacto social de sus herramientas, incluidos sus usos maliciosos.

- Adhesión a prácticas rigurosas de gobernanza y seguridad de datos para priorizar la privacidad y el cumplimiento de las leyes y regulaciones de protección de usuarios.

Por su parte, Microsoft establece seis principios para mejorar la concienciación de su plantilla sobre las cuestiones éticas relacionadas con la IA: equidad, fiabilidad, seguridad y privacidad, inclusividad, transparencia y responsabilidad. De este modo, pretende poner en práctica la gestión de sus políticas mediante la gobernanza, la habilitación y la capacitación para lograr que la aplicación de sus procesos sea responsable. Con el objetivo de minimizar los contenidos estereotipados como un efecto indeseado de su tecnología, así como para no degradar o borrar grupos demográficos identificados y marginados o sobrerrepresentar a otros colectivos, ha establecido procesos para aumentar la diversidad social entre los participantes en la creación de estas herramientas.

Por otro lado, existe un sistema de revisión de los resultados y análisis del sistema productivo desde una perspectiva diversa que incluye miradas de grupos infrarrepresentados, como las mujeres, las personas racializadas o las clases sociales no dominantes. Otras estrategias para detectar y eliminar contenidos abusivos que lleva a cabo Microsoft consiste en potenciar seis áreas: una arquitectura de seguridad sólida en base a estándares responsables de IA y de seguridad; procedencia duradera de los medios y marca de agua para asegurar la trazabilidad de los contenidos mediante los metadatos; luchar contra los abusos y proteger los contenidos para evitar conductas inapropiadas y engañosas; potenciar la colaboración con las autoridades y la sociedad civil; contribuir a la modernización de la legislación; y, por último, concienciar y educar en el uso de la IA para potenciar su visión crítica y su capacidad de análisis sobre los contenidos obtenidos con estas herramientas.

La empresa afirma mantener un proceso de mejora continua sobre sus directrices éticas aplicadas a la IA y en él tienen en cuenta, además de a la propia empresa, a diseñadores y usuarios, por lo que publica documentación y guías de uso. Las recomendaciones para los usuarios se estructuran en base a los tres procesos en los que se aplican sus herramientas. Al comienzo, que tengan claro sus funcionalidades y el margen de error que se ha detectado. Durante la interacción con la IA, que puedan darse cuenta del momento en que la tecnología no está sirviendo a sus intereses para poder interrumpir su uso, así como mostrar información que sea relevante según el contexto social para evitar resultados estereotipados. En el caso de detectar fallos, que se facilite el descarte de información y se ofrezca contenido corregido, así como la posibilidad de matizar y refinar los resultados. A lo largo del tiempo facilitar la personalización y adaptación de la tecnología a las necesidades y usos de cada individuo.

En el caso de Google, la entidad se compromete a identificar efectos injustos derivados de sus tecnologías, especialmente vinculados a aspectos como raza, etnia, género, ingresos económicos, orientación sexual, etc. Para lograrlo ha conformado un equipo central de revisiones éticas antes del lanzamiento de los productos en el que colaboran expertos internos en temas como equidad en el aprendizaje automático, seguridad, privacidad, Derechos Humanos, Ciencias Sociales… Algunas de sus tareas son asesorar a los equipos y actualizar sus informaciones con investigaciones, analizar la escala y alcance de los beneficios o daños derivados de una tecnología, recomendar

evaluaciones técnicas y verificar que no se producen sesgos injustos, y decidir si continúa la línea de desarrollo la tecnología en la que se esté trabajando.

Para Google, los principios de desarrollo de la IA se concretan en:

- Ser socialmente beneficioso.
- Evitar crear o reforzar prejuicios injustos.
- Ser construido y probado de modo seguro.
- Ser responsable ante la población.
- Incorporar principios de diseño de privacidad.
- Mantener altos estándares de excelencia científica.
- Estar disponible para su uso de acuerdo con estos principios.

Por otra parte, se compromete a no diseñar ni implementar IA en las siguientes áreas de aplicación:

- Aquellos que puedan causar daño general.
- Tecnologías destinadas principalmente a causar lesiones.
- Vigilancia que viola normas internacionalmente aceptadas.
- El propósito contraviene el Derecho Internacional y los Derechos Humanos.

En cuanto al proceso de conceptualización y desarrollo, Google plantea una serie de ejes para sus herramientas de IA:

-Diseñar para la responsabilidad: identificando y documentando los daños potenciales para iniciar el desarrollo de los productos de modo que se aborden esos problemas de modo proactivo.

- Realizar pruebas contradictorias: someter a los modelos a pruebas de estrés internas e inclusivas para identificar y mitigar riesgos futuros.

- Comunicar de forma útil, sencilla y transparente: de forma que los usuarios entiendan las funcionalidades y puedan desarrollar usos positivos de las nuevas herramientas.

Anthropic elaboró en 2023 *Claude's Constitution,* una constitución para su sistema de IA, bajo la premisa de que en vez de usar valores determinados implícitamente a través de la retroalimentación humana a gran escala en sus respuestas de IA, ofrece a los modelos lingüísticos valores explícitos determinados por una constitución explícita y detallada. Es lo que denominan 'IA Constitucional' y afirman que es de gran utilidad tanto para mejorar los sistemas como para proteger a sus empleados de contenido dañino y favorecer la transparencia. Estos principios, según la marca, se esfuerzan por no representar

únicamente perspectivas occidentales y se basan en la Declaración de Derechos Humanos de las Naciones Unidas, en buenas prácticas de confianza y seguridad y en principios propuestos por investigadores y otras entidades vinculadas a la IA.

Desde la perspectiva de Anthropic, el sistema de retroalimentación humana sobre los resultados del modelo determina implícitamente los principios y valores que guiaban el comportamiento del modelo, pero esto provoca no sólo una gran inversión de tiempo y recursos, sino que no sea escalable y que los humanos deban interactuar con contenidos perturbadores para moderarlos.

La IA Constitucional, cuyos principios están en continuo desarrollo, se aplica en dos fases: un primer entrenamiento para criticar y revisar sus propias respuestas en base a los principios y algunos ejemplos; y una segunda aplicación basada en aprendizaje por refuerzo con retroalimentación generada por IA, encargada de localizar los resultados más inofensivos.

Por último, la empresa Anthropic también pone el foco en proporcionar instrucción adicional a un modelo de lenguaje sencillo para que se estimulen modelos que reduzcan significativamente los resultados sesgados. Por ello ha desarrollado una biblioteca de *prompts* y un navegador de dilemas éticos para ayudar a los usuarios a abordar problemas complejos desde diferentes perspectivas.

Además, ha abierto el acceso a los *prompts* de los 3 modelos actuales de Claude 3.5: Haiku, Opus y Sonnet. Estos *prompts* de sistema liberados marcan un hito en cuanto a la transparencia de la IA al ofrecer información sobre las instrucciones, límites y estilos que se aplican a esta tecnología. Esta organización también está desarrollando distintas investigaciones sobre la ética de la IA para mitigar sus vulnerabilidades, ya que, entre otras cuestiones, han descubierto que con una batería de preguntas aparentemente inocuas se puede entrenar al sistema para, posteriormente, ofrezca información sobre temas potencialmente dañinos que en principio no pasarían los filtros.

6 DISCUSIÓN Y CONCLUSIONES

En concordancia con investigaciones previas (Cave; Dihal, 2020; Lamensch, 2023) y con las advertencias de organismos internacionales (Equality Now, 2023), los resultados muestran que la IA, de no poner medidas, puede perpetuar estereotipos de raza, género y clase. Esto se debe a que los datos que alimentan estas herramientas reproducen un

contexto en el que las minorías y las mujeres están infrarrepresentadas frente a los colectivos con poder (Doukhan; Poels; Rezgui; Carrive, 2018; Ramasubramanian; Banjo, 2024). Por lo tanto, existe un problema de base en los contenidos que algunas herramientas de IA emplean para obtener información.

Y es que los imaginarios sociales que planteaba Castoriadis (1983) siguen presentes tanto en el sistema mediático como en sus nuevas extensiones. La IA aplicada a las herramientas de comunicación y creación de contenidos, como es el caso de las que aplican IA generativa, constituyen una extensión de la esfera digital. Se trata de un nuevo tipo de herramienta con la que construir contenidos que luego pueden ser empleados en diversos soportes. No obstante, debe mantenerse una perspectiva crítica sobre sus productos para evitar que incidan en sesgos. Esto coincide con los hallazgos del estudio, pues es una de las preocupaciones tanto de las empresas como de las instituciones internacionales.

De este modo, para evitar los puntos ciegos creados por la falta de diversidad que existe en la sociedad actual -y que tiene su reflejo en sus productos culturales e informativos- es necesario, en primer lugar, tener conciencia de esta desigualdad y, en segundo término, tomar medidas para mitigarla.

Los resultados muestran que las principales corporaciones de desarrollo de IA afirman tener presente esta situación y han desarrollado e implementado códigos de conducta y una autorregulación que mejore sus resultados. Esta gobernanza concuerda con las recomendaciones de organismos internacionales (Council of Europe, 2023; UNESCO, 2023) y sirve para regir su comportamiento, al tiempo que los protege de responsabilidades futuras y constituye el pilar de su Responsabilidad Social Corporativa.

Por otro lado, y en coincidencia con estudios previos (Drabiak et al., 2023; Johnson, 2019), la aplicación de la inteligencia artificial y del *Machine Learning* o aprendizaje automático requiere que los desarrolladores demuestren que estas tecnologías funcionan según lo previsto y que se adoptan estrategias para minimizar los riesgos de fallo o sesgo con el objetivo de mantener la ética en sus aplicaciones. Tal y como indican Gray y Witt, los resultados de esta investigación señalan que la aplicación de una ética que asegure la reducción de sesgos no es posible a menos que todas las partes interesadas se impliquen y asuman la responsabilidad en los procesos de creación y personalización de contenidos.

En este contexto, es poco probable que se logre una ética feminista del cuidado de los datos a menos que todas las partes interesadas, incluidas las mujeres, los hombres y

las personas no binarias y transgénero, asuman la responsabilidad de este trabajo tan necesario.

Este trabajo da respuesta a las preguntas de investigación que se planteaba. Estas se centraban en si es posible evitar la reproducción de estereotipos de clase, género y raza en los productos generados a través de la IA (PI1) y, de ser así, en cómo pueden los usuarios detectarlos y minimizar su aparición (PI2). En primer lugar, se puede concluir que evitarlos totalmente aún no es posible debido a que las fuentes de las que se alimenta la IA son la información disponible en la red y lo aprendido de los propios usuarios, por lo que, al menos por el momento, se configurará a su imagen y semejanza. No obstante, sí que es posible y recomendable que los usuarios sean capaces de detectar las representaciones injustas, irreales y estereotipadas de ciertos colectivos, y, en la medida de sus posibilidades, luchar contra ellas y minimizar su generación y difusión.

Las cuatro grandes empresas del sector de la IA (OS1) mantienen que, a través de la autorregulación que han adoptado dentro de sus estrategias de responsabilidad social, están trabajando por generar unos contenidos más igualitarios y menos discriminatorios, además de incorporar a mujeres y miembros de minorías como desarrolladores. También plantean códigos de conductas apropiadas para detectar contenidos inapropiados.

Los códigos de buenas prácticas de instituciones internacionales (OS2) inciden en la necesidad de aumentar la presencia de nuevos perfiles profesionales, así como de incidir en el desarrollo de marcos normativos y en la formación del uso de estas nuevas herramientas.

Por su parte, los principales problemas en la interacción de los usuarios con la IA (OS3) son la falta de capacidad crítica para detectar su presencia, así como la no percepción de los sesgos presentes en los contenidos generados a partir de ella.

De este modo, la investigación alcanza su objetivo general al incorporar recomendaciones para que los usuarios hagan un uso ético de esta nueva tecnología. Para evitar caer en estereotipos de clase, raza y/o género deben replantearse la información obtenida de las herramientas en cada fase del proceso; adecuar el uso de los *prompts* para evitar incurrir en sesgos; y reportar contenidos inadecuados, poco precisos o arquetípicos; trabajar de forma cíclica todo el proceso para contribuir a un entrenamiento de la IA que genere contenidos más igualitarios, entre otras cuestiones.

El aporte principal de este trabajo consiste en ofrecer una perspectiva general de los procesos que pueden influir en la generación mediante herramientas de IA de contenidos

estereotipados que no aporten una representación realista en cuanto a raza, género y clase. Sus hallazgos son útiles para fomentar la perspectiva crítica y poner el foco en el papel que los individuos tienen en un proceso en el que el uso que se haga de las herramientas resulta fundamental.

La lucha contra la desigualdad, los sesgos, los estereotipos y los prejuicios constituye una elección política y social, además de ser una obligación ética para las empresas. Las principales corporaciones ya han reconocido la falta de neutralidad de los sistemas de IA y afirman estar poniendo medidas para no reproducir y amplificar la desigualdad estructural y los mecanismos que facilitan la exclusión de personas y colectivos. Como recomendaciones para aumentar la equidad en el mercado de la IA se pueden distinguir dos aspectos:

- medidas a adoptar durante el desarrollo de herramientas: disponer de equipos de perfiles variados, con presencia integrada de minorías y mujeres que puedan detectar sesgos específicos; dotarse de códigos éticos y medidas de autorregulación -parece una buena iniciativa la idea de crear una Constitución desglosada acorde a los valores de la marca- que especifiquen las líneas maestras a seguir; facilitar el uso de *prompts*; promover la transparencia y la reparación tanto de contenidos como hacia personas afectadas; y fomentar la colaboración multidisciplinar para potenciar los buenos usos de esta tecnología y detectar las desviaciones para erradicar los usos negativos.
- medidas respecto al uso de las herramientas: formar a los usuarios para potenciar el espíritu crítico; promover la evaluación de los productos obtenidos a través de la IA; y facilitar los procesos de empleo de las herramientas, adaptándolas a las necesidades de los usuarios y facilitando el reporte de resultados inadecuados.

Con todo lo descrito, y pese a las limitaciones de esta investigación, que son las propias de un trabajo exploratorio que se centra en un área de estudio reciente y de rápida evolución, puede ser de utilidad tanto para usuarios como para desarrolladores.

No obstante, se trata de un área de estudio amplia y dinámica, que puede afrontarse desde nuevas líneas de investigación, como los usos y gratificaciones en relación con estas herramientas, los diferentes tipos de contenido que pueden generarse mediante esta tecnología y el papel que juegan los usuarios en cada uno de ellos.

REFERENCIAS

ADAMS, Rachel; LOIDEÁIN, Nora Ni. Addressing indirect discrimination and gender stereotypes in AI virtual personal assistants: the role of international human rights law. **Cambridge International Law Journal**, 8(2), 241-257, 2019. Disponible en: http://dx.doi.org/10.2139/ssrn.3392243 Acceso en: 1 oct. 2024.

ADCOCK, Charlotte. The Politician, The Wife, The Citizen, and her Newspaper, **Feminist Media Studies**, 10(2), 135-159, 2010. Disponible en: https://doi.org/10.1080/14680771003672254 Acceso en: 12 oct. 2024.

AHMED, Jashim Uddin. Documentary research method: New dimensions. **Indus Journal of Management & Social Sciences**, 4(1), 1-14, 2010. Disponible en: RePEc:iih:journl:v:4:y:2010:i:1:p:1-14 Acceso en: 10 oct. 2024.

AHN, Jungyong; KIM, Jungwon; SUNG, Yongjun. The effect of gender stereotypes on artificial intelligence recommendations. **Journal of Business Research**, 141, 50-59, 2022. Disponível em: https://doi.org/10.1016/j.jbusres.2021.12.007 Acceso en: 2 oct. 2024.

ARJONA-MARTÍN, José-Borja; MÉNDIZ-NOGUERO, Alfonso; VICTORIA-MÁS, Juan Salvador. Virality as a paradigm of digital communication. Review of the concept and update of the theoretical framework. **Profesional de la información,** 29(6), e290607, 2020. Disponible en: https://doi.org/10.3145/epi.2020.nov.07 Acceso en: 25 sep. 2024.

BARRIENTOS-BÁEZ, Almudena; PIÑEIRO OTERO, María Teresa; PORTO RENÓ, Denis. Imágenes falsas, efectos reales. Deepfakes como manifestaciones de la violencia política de género. **Revista Latina de Comunicación Social**, n. 82, p. 1–30, 2024. DOI: 10.4185/rlcs-2024-2278. Disponible en: https://nuevaepoca.revistalatinacs.org/index.php/revista/article/view/2278 Acceso en: 13 dec. 2024.

BAUTISTA, Pavel S.; CABEZUELO-LORENZO, Francisco; DE LA CASA, José María H. Instagram como herramienta digital para la comunicación y divulgación científica: el caso mexicano de@ pictoline**. Chasqui: Revista latinoamericana de comunicación,** (147), 143-162, 2021. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=8093846 Acceso en: 3 oct, 2024.

BAZZANO, Manu (2023). Everybody wants to be a manager: On masculinity, microfascism and the manosphere. In **Psychotherapy and Unstable Notions of Masculinity** (pp. 21-34). Londres: Routledge. 2023.

BLIGH, Michelle C., SCHLEHOFER, Michelle M., CASAD, Bettina J. y GAFFNEY, Amber. M. (2012). Competent enough, but would you vote for her? Gender stereotypes and media influences on perceptions of women politicians. **Journal of Applied Social Psychology**, 42(3), 560-597. Disponible en: https://doi.org/10.1111/j.1559-1816.2011.00781.x Acceso en: 12 oct. 2024.

BRUNO, Marco. Media representations of immigrants in Italy: framing real and symbolic borders. **REMHU: Revista Interdisciplinar da Mobilidade Humana**, (24), 45-58, 2016. Disponible en: https://doi.org/10.1590/1980-85852503880004604. Acceso en: 12 dic. 2024.

BROOKS, Dwight., & HÉBERT, Lisa. Gender, race, and media representation. **Handbook of gender and communication**, 16, 297-317, 2006. Disponible en: https://doi.org/10.4135/9781412976053.n16 Acceso en: 25 sep. 2024.

VERGÉS, Núria; ALFAMA, Eva; Cruells, Eva. Violències masclistes digitals: implicacions del seu abordatge en el marc dels circuits de violència masclista. **Idees: Revista de temes contemporanis**, (59), 8, 2022.

BYERLY, Carolyn M.; ROSS, Karen. **Women and media: A critical introduction**. Hoboken: Blackwell. 2008.

CARTY, Victoria. **Social movements and new technology**. Boulder: Westview Press. 2015.

CASTEJÓN, María. Mujeres y cine: las fuentes cinematográficas para el avance de la historia de las mujeres. **Berceo,** (147), 303-327, 2004. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=1387383 Acceso en: 2 oct. 2024.

COLELLA, Francesca; GIANTURCO, Giovanna. L'idea sociale delle migrazioni nella società contemporanea. **Sociologia e ricerca sociale**, (123), pp. 5-18, 2020. Disponible en: http://digital.casalini.it/10.3280/SR2020-123001. Acceso en: 12 dic. 2024.

CABALLERO, Jimena Yisel. LA APROPIACIÓN DE LAS MUJERES CON INTELIGENCIA ARTIFICIAL: DEL CUERPO AL AMOR. **Inteligencia Artificial y Comunicación**, 85-102. Aveiro: Ria Editorial. 2024.

CABRERA, Daniel. **Imaginario social, comunicación e identidad colectiva. Diálogo Comunicación y diversidad cultural**, Forum Barcelona, Institut de la Comunicació-Universitat Autònoma de Barcelona, 1-15, 2004. Disponible en: https://www.researchgate.net/publication/242731193_Imaginario_social_comunicacion_e_identidad_colectiva Acceso en: 29 sep. 2024.

CASTORIADIS, Cornelius. **La institución imaginaria de la sociedad**. Barcelona: Tusquets. 1983.

CAVE, Stephen; DIHAL, Kantal. The whiteness of AI**. Philosophy & Technology**, 33(4), 685-703, 2020. Disponible en: https://doi.org/10.1007/s13347-020-00415-6 Acceso en: 12 oct. 2024.

CHERYAN, Sapna; PLAUT, Victoria C.; HANDRON, Caitlin; HUDSON, Lauren. The stereotypical computer scientist: Gendered media representations as a barrier to inclusion for women**. Sex roles,** 69(1), 58-71, 2013. Disponible en: https://doi.org/10.1007/s11199-013-0296-x Acceso en: 18 sep. 2024

CRAIUT, Miruna-Valeria; IANCU, Ioana. Is technology gender neutral? A systematic literature review on gender stereotypes attached to artificial intelligence. **Human Technology**, 18(3), 297-315, 2022. Disponible en: https://doi.org/10.14254/1795-6889.2022.18-3.6 Acceso em: 13 oct. 2024.

COUNCIL OF EUROPE (2023). **Women in Media and artificial intelligence**. Council of Europe. Disponible en: https://www.coe.int/en/web/genderequality/women-in-media Acceso en: 13 octubre, 2024.

CUENCA, Nerea. Cambio de paradigma narrativo en la Era Postfeminista: estudio de caso de Un lugar para soñar (Sue Teney, 2019-2020). **Miguel Hernández Communication Journal**, 13(1), 125-144, 2022. Disponible en: https://doi.org/10.21134/mhjournal.v13i.1375 Acceso en: 3 oct. 2024.

CURRAN, James. Rethinking the media as a public sphere. In P. Dahlgren, & C. Sparks (Eds.) **Communication and citizenship** (pp. 27-57). Londres: Routledge. 2005.

DAVIS, Angela. Rape, racism and the capitalist setting. **The Black Scholar**, 12(6), 39-45, 1981. Disponible en: https://doi.org/10.1080/00064246.1981.11414219. Acceso en: 12 dic. 2024.

DAVIS, Angela. **Women, race, & class**. Nueva York: Vintage. 2011

DAVIS, Angela. **Angela Davis: autobiografía**. Madrid: Capitán Swing Libros. 2017.

DRABIAK, Katherine; KYZER, Skylar; NEMOV, Valerie; EL NAQA, Issam. AI and machine learning ethics, law, diversity, and global impact. **The British journal of radiology**, 96(1150), 20220934, 2023. Disponible en: https://doi.org/10.1259/bjr.20220934 Acceso en: 12 dic. 2024.

DUBBER, Markus D.; PASQUALE, Frank; DAS, Sunit (Eds.). **The Oxford handbook of ethics of AI.** Oxford: Oxford Handbooks. 2020. Disponible en: https://doi.org/10.1093/oxfordhb/9780190067397.001.0001 Acceso em: 3 oct. 2024.

DUKES, Kristin Nicole; GAITHER, Sarah E. Black racial stereotypes and victim blaming: Implications for media coverage and criminal proceedings in cases of police violence against racial and ethnic minorities. **Journal of Social Issues**, 73(4), 789-807, 2017. Disponible en: https://doi.org/10.1111/josi.12248 Acceso en: 17 sep. 2024.

DURMUS, Koçak. The Impact of Cyber Culture on New Media Consumers. In: DAVID, Shibin, R.S. ANAND; V. JEYAKRISHNAN; M. NIRANJANAMURTHY, **Security Issues and Privacy Concerns in Industry 4.0 Applications**, (229-247), 2021. Scrivener Publishing LLC. Disponible en: https://doi.org/10.1002/9781119776529.ch12 Acceso en: 10 oct. 2024.

DOUKHAN, David; POELS, Géraldine; REZGUI, Zohra; CARRIVE, Jean. Describing Gender Equality in French Audiovisual Streams with a Deep Learning Approach. **VIEW Journal of European Television History and Culture**, 7(14), 103–122, 2018. Disponible en: https://doi.org/10.25969/mediarep/14757 Acceso en: 3 oct. 2024.

EKSI, Sena. Digital Populism: The Internet and the Rise of Right-Wing Populism. **ECPS**. 2022. Disponible en: https://www. populismstudies.org/digital-populism-the-internet-and-the-rise-of-right-wing-populism/. Acceso en: 13 dic. 2024.

EUROPA PRESS PORTALTIC. **OpenAI, Google, Microsoft y Anthropic se unen para autorregular el desarrollo de los modelos de IA más sofisticados**. EUROPA PRESS PORTALTIC. 2023. Disponible en: https://www.europapress.es/portaltic/sector/noticia-openai-google-microsoft-anthropic-unen-autorregular-desarrollo-modelos-ia-mas-sofisticados-20230726143239.html Acceso en: 12 oct. 2024.

FOX, Stephen. Domesticating artificial intelligence: Expanding human self-expression through applications of artificial intelligence in prosumption. **Journal of Consumer Culture**, 18(1), 169-183, 2018. Disponible en: https://doi.org/10.1177/1469540516659126 Acceso en: 14 oct. 2024.

FREIXAS, Anna. Entre el mandato y el deseo: el proceso de adquisición de la identidad sexual y de género". En FLECHA, Consuelo y NÚÑEZ, Marina (eds.). **La Educación de las Mujeres: Nuevas perspectivas** (pp. 23-32). Sevilla: Secretariado de publicaciones de la Universidad de Sevilla. 2000.

GING, Debbie. Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. **Men and Masculinities**, 22(4), 638-657, 2019. Disponible en: https://doi.org/10.1177/1097184X17706401. Acceso en: 13 dic. 2024.

GONZÁLEZ, Blanca. Los estereotipos como factor de socialización en el género. **Comunicar**, (12), 79-88, 1999. Disponible en: https://www.redalyc.org/pdf/158/15801212.pdf Acceso en: 5 oct. 2024.

GRAY, Joanne; WITT, Alice. A feminist data ethics of care for machine learning: The what, why, who and how. **First Monday**, [S. l.], v. 26, n. 12, 2021. DOI: 10.5210/fm.v26i12.11833. Disponible en: https://firstmonday.org/ojs/index.php/fm/article/view/11833 . Acceso en: 13 dic. 2024.

HESMONDHALGH, David. The media's failure to represent the working class: Explanations from media production and beyond 1. In DEERY, June y PRESS, Andrea, **Media and class** (pp. 21-37). Londres: Routledge. 2017. Disponible en: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315387987-2/media-failure-represent-working-class-explanations-media-production-beyond-1-david-hesmondhalgh Acceso en: 2 oct. 2024.

HORTA, Manoel; BLACKBURN, Jeremy; BRADLYN, Barry; DE CRISTOFARO, Emiliano; STRINGHINI, Gianluca; LONG, Summer; GREENBERG, Stephanie; ZANNETTOU, Savvas. The Evolution of the Manosphere across the Web. **Proceedings of the International AAAI Conference on Web and Social Media**, [S. l.], v. 15, n. 1, p. 196-207, 2021. DOI: 10.1609/icwsm.v15i1.18053. Disponible en: https://ojs.aaai.org/index.php/ICWSM/article/view/18053. Acceso en: 13 dic. 2024.

JOHNSON, Sandra L. J. AI, Machine Learning, and Ethics in Health Care. **Journal of Legal Medicine**, 39(4), 427–441, 2019. Disponible en: https://doi.org/10.1080/01947648.2019.1690604. Acceso en: 13 dic. 2024.

JOYCE, Mary C. **Digital activism decoded: The new mechanics of change**. Santa Cruz de Tenerife: IDEA. 2010.

JENKINS, Henry; FORD, Sam; GREEN, Joshua. **Spreadable media: Creating value and meaning in a networked culture**. 2013. Nueva York: NYU Press. Disponible en: http://www.jstor.org/stable/j.ctt9qfk6w Acceso en: 28 sep. 2024.

KENDALL, Diana E. **Framing class: Media representations of wealth and poverty in America**. Nueva York: Rowman & Littlefield. 2011.

KIM, Juran, KANG; Seungmook; LEE, Ki Hoon. Evolution of digital marketing communication: Bibliometric analysis and network visualization from key articles. **Journal of Business Research**, 130, 552-563, 2021. Disponible en: https://doi.org/10.1016/j.jbusres.2019.09.043 Acceso en: 12 oct. 2024.

KOTLER, Philip. The Prosumer Movement. In: Blättel-Mink, B., Hellmann, KU. (eds) **Prosumer Revisited**. Wiesbaden: VS Verlag für Sozialwissenschaften. 2010. Disponible en: https://doi.org/10.1007/978-3-531-91998-0_2 Acceso en: 12 oct. 2024.

LAGARDE, Marcela. **Identidad genérica y feminismo**. Sevilla: Instituto Andaluz de la Mujer. 1998.

LAMENSCH, Marie. **Generative AI Tools Are Perpetuating Harmful Gender Stereotypes**. Waterloo: Centre for International Governance Innovation. 2023. Disponible en: https://www.cigionline.org/articles/generative-ai-tools-are-perpetuating-harmful-gender-stereotypes/ Acceso en: 28 sep. 2024.

LEI, Rayan F.; BODENHAUSEN, Galen V. Racial assumptions color the mental representation of social class. **Frontiers in psychology,** 8, 519, 2017. Disponible en: https://doi.org/10.3389/fpsyg.2017.00519 Acceso en: 12 oct. 2024.

LEVY, Pierry. **Cibercultura**. Sao Paulo: Editora 34. 2010.

LÓPEZ, Clara; VILASECA, Carol, & SERRANO, Jazmín Mariana.Interseccionalidad: la discriminación múltiple desde una perspectiva de género. **Revista Crítica de la Historia de las Relaciones Laborales y de la Política Social,** (14), 71-81, 2022. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=8928082 Acceso en: 14 diciembre, 2024.

MANASI, Ardra; PANCHANADESWARAN, Subadra; SOURS, Emily. **Addressing Gender Bias to Achieve Ethical AI.** The Global Observatory. 2023. Disponible en: https://theglobalobservatory.org/2023/03/gender-bias-ethical-artificial-intelligence/ Acceso en: 28 sep. 2024.

MARWICK, Alice; CAPLAN, Robyn. Drinking male tears: Language, the manosphere, and networked harassment. **Feminist media studies**, 18(4), 543-559, 2018. https://doi.org/10.1080/14680777.2018.1450568

MASTRO, Dana. Effects of racial and ethnic stereotyping. In BRYANT, Jennings y OLIVER, Mary Beth, **Media effects** (pp. 341-357). Londres: Routledge. 2009.

MCQUILLAN, Dan. **Resisting AI: an anti-fascist approach to artificial intelligence.** Policy Press. 2022.

MONEDERO, Carmen; IMPELLUSO, Pablo. Empoderamiento femenino y cultura urbana en las series Bravas y La Reina del Flow, **Miguel Hernández Communication Journal**, 12(2), 587-609, 2021. Disponible en: https://doi.org/10.21134/mhjournal.v12i.1329 Acceso en: 28 sep. 2024.

MUNT, Sally (Ed.). **Cultural studies and the working class**. Londres: A&C Black. 2000.

MUSSA, I. (2019). Ódio ao jogo: cripto-fascismo e comunicação anti-lúdica na cultura dos videogames. **Logos**, 26(2), 57-71.

PARRA-VALERO, Pablo; OLIVEIRA, Lídia. Fake news: una revisión sistemática de la literatura, **Observatorio (OBS*)**, 12(5): 54-78, 2018. Disponible en: https://hdl.handle.net/20.500.14352/99732 Acceso en: 12 oct. 2024.

TEIXEIRA, Ana. Capitalism with a Transhuman Face: The Afterlife of Fascism and the Digital Frontier. **Third Text**, 33(3), 315–336, 2019. Disponible en: https://doi.org/10.1080/09528822.2019.1625638. Acceso en: 13 dic., 2024

PIÑEIRO-OTERO, Teresa; MARTÍNEZ-ROLÁN, Xabier. Eso no me lo dices en la calle. Análisis del discurso del odio contra las mujeres en Twitter, **Profesional de la información**, 30(5), e300402, 2021. Disponible en: https://doi.org/10.3145/epi.2021.sep.02 Acceso en: 12 oct. 2024.

RAMASUBRAMANIAN, Srividya; BANJO, Omotayo O. (Eds.). **The Oxford Handbook of Media and Social Justice**. Oxford: Oxford University Press. 2024. Disponible en: https://doi.org/10.1093/oxfordhb/9780197744345.001.0001 Acceso en: 28 sep. 2024.

ROSENTHAL, Lisa; LOBEL, Marci. Gendered racism and the sexual and reproductive health of Black and Latina Women. **Ethnicity & Health**, 25(3), 367–392, 2018. Disponible en: https://doi.org/10.1080/13557858.2018.1439896. Acceso en: 13 dic. 2024.

RUFFINELLI, Jorge. Ellas lo hacen mejor (Un cine nuevo: el de mujeres), **Cinémas d'Amérique latine**, (22), 80-89, 2014. Disponible en: https://doi.org/10.4000/cinelatino.785 Acceso en: 12 oct. 2024.

RUGGIERO, Thomas E. Uses and gratifications theory in the 21st century. **Mass communication & society**, 3(1), 3-37, 2000. Disponible en: https://doi.org/10.1207/S15327825MCS0301_02. Acceso en: 13 dic. 2024.

SCHARRER, Erica; RAMASUBRAMANIAN, Srividya. Intervening in the media's influence on stereotypes of race and ethnicity: The role of media literacy education**, Journal of Social Issues**, 71(1), 171-185, 2015. Disponible en: https://doi.org/10.1111/josi.12103 Acceso en: 12 oct. 2024.

SOLA-MORALES, Salomé. Imaginarios sociales, procesos de identificación y comunicación mediática, **Prisma**, (25), 3-22, 2014. Disponible en: http://193.137.34.194/index.php/prismacom/article/view/1866/1699 Acceso en: 28 sep. 2024.

SOSA-VALCARCEL, Aimiris; GALARZA-FERNÁNDEZ, Emelina; CASTRO-MARTINEZ, Andrea. Acción colectiva ciberactivista de "Las periodistas paramos" para la huelga feminista del 8M en España, **Comunicación y Sociedad**, e7287, 2019. Disponible en: https://doi.org/10.32870/cys.v2019i0.7287. Acceso en: 28 sep. 2024.

SPATES, Kamesha; EVANS, Na'tasha; JAMES, Tierra Akilah; MARTINEZ, Karen. Gendered Racism in the Lives of Black Women: A Qualitative Exploration. **Journal of Black Psychology**, 46(8), 583-606, 2020. Disponible en: https://doi.org/10.1177/0095798420962257. Acceso en: 14 dic. 2024.

TANJITPIYANOND, Porntida; JETTEN, Jolanda; PETERS, Kim. How economic inequality shapes social class stereotyping, **Journal of Experimental Social Psychology**, 98, 104248, 2022. Disponible en: https://doi.org/10.1016/j.jesp.2021.104248 Acceso en: 12 oct. 2024.

THORNHAM, Sue. **Women, feminism and media**. Edimburgo: Edinburgh University Press. 2007

TORRES-MARTÍN, José Luis; CASTRO-MARTÍNEZ, Andrea; DÍAZ- MORILLA; Pablo; PÉREZ-ORDÓÑEZ, Cristina.Mujeres directivas y creadoras en el audiovisual. Análisis de las series de ficción españolas presentes en los catálogos de Amazon Prime Video, Movistar+ y Netflix (2019-2021). **Perspectivas De La Comunicación**, 15(2), 217–248, 2022. Disponible en: https://doi.org/10.56754/0718-4867.1502.217 Acceso en: 28 sep. 2024.

TORRES-MARTÍN, José Luis; CASTRO-MARTÍNEZ, Andrea; PÉREZ ORDÓÑEZ, Cristina. Las mujeres directivas en el audiovisual español en la era de las plataformas de contenidos bajo demanda. Una revisión documental y bibliográfica, en Izquierdo, J: **Mujeres en streaming. Especialización, liderazgo y representación** (31-53). Madrid: Fragua. 2023.

TUKACHINSKY, Riva; MASTRO, Dana; YARCHI, Mora. The Effect of Prime Time Television Ethnic/Racial Stereotypes on Latino and Black Americans: A Longitudinal National Level Study, **Journal of Broadcasting & Electronic Media**, 61(3), 538–556, 2017. Disponible en: https://doi.org/10.1080/08838151.2017.1344669 Acceso en: 28 sep. 2024.

TURNER, Fred. **Aux sources de l'utopie numérique: de la contre-culture à la cyberculture, Stewart Brand, un homme d'influence**. Caen: C & F Éditions. 2021.

UNESCO. **Intelligence artificielle : l'UNESCO lance le Réseau des femmes pour une IA éthique, plateforme d'expertes pour faire progresser l'égalité des genres.** Communiqué de presse, UNESCO, 30 avril 2023. Disponible en: https://www.unesco.org/fr/articles/intelligence-artificielle-lunesco-lance-le-reseau-des-femmes-pour-une-ia-ethique-plateforme Acceso en: 12 oct. 2024.

VAN DIJK, Teun A.. **Discurso y poder**. Barcelona: Editorial Gedisa. 2011

WEST, Sarah M.; WHITTAKER, Meredith; CRAWFORD, Kate. **Discriminating systems. AI Now**, 1-33, 2019. Disponible en: https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2 Acceso en: 12 oct. 2024.

NOTAS

USO DE INTELIGENCIA ARTIFICIAL
No aplica.

FINANCIAMIENTO
No aplica.

CONSENTIMIENTO PARA EL USO DE LA IMAGEN
No aplica.

APROBACIÓN DEL COMITÉ DE ÉTICA EN LA INVESTIGACIÓN
No aplica.

CONFLICTO DE INTERÉS
Las autorías declaran que no existen conflictos de intereses.

DISPONIBILIDAD DE DATOS DE INVESTIGACIÓN Y OTROS MATERIALES
URL o DOI 1: https://www.coe.int/en/web/genderequality/-/tackling-bias-in-artificial-intelligence- systems-to-promote-equality-new-study-published
URL o DOI 2: https://rm.coe.int/prems-112923-gbr-2530-etude-sur-l-impact-de-ai-web-a5-1-2788- 3289-7544/1680ac7936
URL o DOI 3: https://www.coe.int/en/web/genderequality/committee-of-experts-on-artificial- intelligence-equality-and-discrimination-gec/adi-ai-
URL o DOI 4: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf
URL o DOI 5: https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial- intelligence
URL o DOI 6: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics URL o DOI 7: https://unesdoc.unesco.org/ark:/48223/pf0000381137
URL o DOI 8: https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming- evidence-regressive-gender-stereotypes
URL o DOI 9: https://unesdoc.unesco.org/ark:/48223/pf0000388971
URL o DOI 10: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases URL o DOI 11: https://docs.anthropic.com/es/prompt-library/ethical-dilemma-navigator
URL o DOI 12: https://www.anthropic.com/responsible-disclosure-policy URL o DOI 13: https://www.anthropic.com/company
URL o DOI 14: https://www.anthropic.com/news/announcing-our-updated-responsible-scaling-policy URL o DOI 15: https://openai.com/index/how-should-ai-systems-behave/
URL o DOI 16: https://openai.com/policies/supplier-code/ URL o DOI 17: https://openai.com/safety/
URL o DOI 18: https://openai.com/index/our-approach-to-ai-safety/ URL o DOI 19: https://openai.com/security-and-privacy/
URL o DOI 20: https://openai.com/policies/
URL o DOI 21: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1l5BO
URL o DOI 22: https://blogs.microsoft.com/on-the-issues/2024/07/30/protecting-the-public-from- abusive-ai-generated-content/
URL o DOI 23: https://blogs.microsoft.com/on-the-issues/2024/07/05/combatting-ai-deepfakes-our- participation-in-the-2024-political-conventions/
URL o DOI 24: https://www.microsoft.com/en-us/haxtoolkit/workbook/?culture=en-us&country=us URL o DOI 25: https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/

URL o DOI 26: https://www.microsoft.com/en-us/haxtoolkit/design-library-overview/ URL o DOI 27: https://www.microsoft.com/en-us/haxtoolkit/playbook/
URL o DOI 28: https://unlocked.microsoft.com/responsible-ai/ URL o DOI 29: https://www.microsoft.com/en-us/ai/responsible-ai

URL o DOI 30: https://learn.microsoft.com/es-es/azure/cloud-adoption- framework/strategy/responsible-ai
URL o DOI 31: https://ai.google/responsibility/principles/
URL o DOI 32: https://ai.google/responsibility/responsible-ai-practices/ URL o DOI 33: https://ai.google/responsibility/responsible-ai-practices/
URL o DOI 34: https://research.google/blog/google-research-2022-beyond-responsible-ai/?_gl=1*vamj79*_ga*MTk4MTc0ODYwMC4xNzI4OTIwNDMw*_ga_KFG60X3H7K*MTcyOTA3OTQ4OC4yLjEuMTcyOTA3OTU4OC4wLjAuMA..
URL o DOI 35: https://cloud.google.com/responsible-ai
URL o DOI 36: https://developers.google.com/machine-learning/resources/intro-responsible- ai?hl=es-419
URL o DOI 37: https://publicpolicy.google/responsible-ai/
URL o DOI 38: https://www.anthropic.com/news/claudes-constitution

EDITORIAL
Universidade Federal de Santa Catarina. Las ideas expresadas en este artículo son responsabilidad de las personas autoras y no necesariamente representan la opinión de los editores o de la universidad.

EDITORES
Edgar Bisset Alvarez, Ana Clara Cândido, Patrícia Neubert, Genilson Geraldo, Jônatas Edison da Silva, Mayara Madeira Trevisol, Edna Karina da Silva Lira e Luan Soares Silva.

HISTORIA
Recebido em: 16-10-2025 – Aprovado em: 30-12-2024 – Publicado em: 14-03-2025