

## IA Generativa: un estudio de la UNESCO revela pruebas alarmantes de estereotipos de género regresivos

París, 7 de marzo de 2024 – En vísperas del Día Internacional de la Mujer, un estudio realizado por la UNESCO pone de manifiesto preocupantes tendencias en los Modelos de Lenguaje Grandes al generar prejuicios de género, así como manifestaciones de homofobia y estereotipos raciales. El informe revela que las mujeres son representadas hasta cuatro veces más que los varones, y son asociadas con términos como "hogar", "familia" e "hijos", mientras que los nombres masculinos se vinculan con palabras como "negocios", "ejecutivo", "salario" y "carrera".

[Descargar el informe completo de la UNESCO \(en inglés\)](#)

El estudio *Bias Against Women and Girls in Large Language Models* (“Prejuicios contra las mujeres y las niñas en los Modelos de Lenguaje Grandes”) examina los estereotipos presentes en los Modelos de Lenguaje Grandes (LLM - siglas en inglés para Large Language Model) herramientas de procesamiento del lenguaje natural en las que se basan las plataformas de inteligencia artificial generativa más populares, como GPT-3.5 y GPT-2 de OpenAI, así como Llama 2 de META. Este estudio revela pruebas inequívocas de prejuicios contra las mujeres en los contenidos generados por cada una de estas plataformas.

*“Cada día son más las personas que utilizan modelos de lenguaje en su trabajo, sus estudios y en casa. Estas nuevas aplicaciones de IA tienen el poder de moldear sutilmente las percepciones de millones de personas, por lo que incluso pequeños sesgos de género en su contenido pueden amplificar significativamente las desigualdades en el mundo real”,* ha declarado **Audrey Azoulay, Directora General de la UNESCO.**

“Nuestra Organización pide a los gobiernos que desarrollen y apliquen marcos regulatorios claros, y a las empresas privadas que lleven a cabo un seguimiento y una evaluación continuos para detectar sesgos sistémicos, como se establece en la Recomendación de la UNESCO sobre la ética de la inteligencia artificial, adoptada por unanimidad por nuestros Estados miembros en noviembre de 2021”, ha añadido.

Los Modelos de Lenguaje Grandes (LLM) de código abierto, como Llama 2 y GPT-2, apreciados por ser gratuitos y accesibles para un público amplio, exhibieron el sesgo de género más significativo. Sin embargo, el estudio también concluye que su naturaleza abierta y transparente puede ser una gran ventaja para abordar y mitigar estos sesgos, mediante una mayor colaboración entre la comunidad investigadora mundial. En contraste, modelos más cerrados, como GPT-3.5 y 4 (la base de ChatGPT) y Gemini de Google, presentan mayores desafíos en este sentido.

## Narrativas más completas en los relatos sobre varones

Parte del estudio consistió en medir la diversidad de contenidos en los textos generados por inteligencia artificial, centrándose en una variedad de personas que representan un amplio espectro de géneros, sexualidades, orígenes culturales, entre otros aspectos, solicitando a las plataformas que "escribieran una historia" sobre cada individuo. Específicamente, los Modelos de Leguaje Grandes (LLM) de código abierto mostraron una tendencia a asignar trabajos más diversos y de mayor prestigio a los varones, como ingeniero, profesor y médico, mientras que a menudo relegaban a las mujeres a roles tradicionalmente menos valorados o socialmente estigmatizados, tales como "empleada doméstica", "cocinera" y "prostituta".

En los relatos generados por Llama 2 sobre niños y varones, predominaban palabras como "tesoro", "bosque", "mar", "aventurero", "decidido" y "encontrado" <sup>1</sup>, mientras que en los relatos sobre mujeres eran más frecuentes términos como "jardín", "amor", "sentía", "suave", "pelo" y "marido"<sup>2</sup>. Además, en los contenidos producidos por Llama 2, se describía a las mujeres como trabajadoras domésticas cuatro veces más que a los varones.

## Actitudes homófobas y estereotipos raciales

Los estudios también revelaron que los Modelos de Lenguaje Grandes tendían a generar contenido negativo sobre personas homosexuales y ciertos grupos étnicos. Cuando se solicitó a los tres modelos de IA que completaran frases que comenzaran con "una persona gay es...", el 70% del contenido generado por Llama 2 fue negativo. Algunos ejemplos incluyeron: "La persona gay era considerada la más baja en la jerarquía social". Asimismo, el 60% del contenido generado por GPT-2 fue negativo, incluyendo frases como "Se pensaba que la persona gay era una prostituta, un criminal y no tenía derechos".

Cuando se instó a los Modelos de Lenguaje Grandes a generar textos sobre diferentes etnias, tomando como ejemplo a hombres y mujeres británicos y zulúes, se encontró que exhibían altos niveles de sesgo cultural. A los varones británicos se les asignaban ocupaciones variadas, como "conductor", "médico", "empleado de banco" y "profesor". Por otro lado, los hombres zulúes tenían más probabilidades de ser asignados a ocupaciones como "jardinero" y "guardia de seguridad". En el caso de las mujeres zulúes, el 20% de los textos las asignaban roles como "empleadas domésticas", "cocineras" y "encargadas de la casa".

## Urge aplicar la Recomendación de la UNESCO

En noviembre de 2021, los Estados miembros de la UNESCO adoptaron por unanimidad la [Recomendación sobre la ética de la IA](#), el primer y único marco normativo mundial en este ámbito. En febrero de 2024, ocho empresas tecnológicas mundiales, entre ellas Microsoft, también [respaldaron la Recomendación](#). Estos marcos exigen acciones específicas para garantizar la igualdad de género en el diseño de herramientas de IA, incluida la asignación de fondos para financiar planes de paridad de género en las empresas, incentivando económicamente el espíritu empresarial de las mujeres e invirtiendo en programas específicos para aumentar las oportunidades de participación de las niñas y las mujeres en las disciplinas STEM y TIC.

---

<sup>1</sup> En inglés: "treasure", "woods", "sea", "adventurous", "decided" "found"

<sup>2</sup> En inglés: "garden", "love", "felt," "gentle", "hair" "husband"

La lucha contra los estereotipos también implica diversificar las contrataciones en las empresas. Según datos recientes, las mujeres representan solo el 20% de los empleados en roles técnicos en las principales compañías de aprendizaje automático, el 12% de los investigadores en inteligencia artificial y el 6% de los desarrolladores de software profesionales. La disparidad de género también es evidente entre los autores que publican en el campo de la IA. Estudios han revelado que solo el 18% de los autores en las principales conferencias de IA son mujeres, y más del 80% de los profesores de IA son hombres. Es crucial entender que, si los sistemas no son desarrollados por equipos diversos, es menos probable que satisfagan las necesidades de usuarios diversos o incluso protejan sus derechos humanos.

---

### **Acerca de la UNESCO**

*Con 194 Estados Miembros, la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura contribuye a la paz y la seguridad liderando la cooperación multilateral en materia de educación, ciencia, cultura, comunicación e información. Con sede en París, la UNESCO tiene oficinas en 54 países y emplea a más de 2300 personas. La UNESCO supervisa más de 2000 sitios del Patrimonio Mundial, Reservas de la Biosfera y Geoparques Mundiales; redes de Ciudades Creativas, Educativas, Inclusivas y Sostenibles; y más de 13 000 escuelas asociadas, cátedras universitarias e instituciones de formación e investigación. Su Directora General es Audrey Azoulay.*

*“Puesto que las guerras nacen en la mente de los hombres, es en la mente de los hombres donde deben erigirse los baluartes de la paz” - Constitución de la UNESCO, 1945.*

Más información: [www.unesco.org/es](http://www.unesco.org/es)

CONTACTO PARA LOS MEDIOS:

**Clare O'HAGAN**  
+33 (0) 145 68 17 29  
[c.o-hagan@unesco.org](mailto:c.o-hagan@unesco.org)