FEMINISMO Y TECNOLOGÍA: EXPLORANDO LOS SESGOS DE GÉNERO EN LA IA GENERATIVA A TRAVÉS DEL PROYECTO LENA

MARÍA ASUNCIÓN VICENTE RIPOLL
CÉSAR FERNÁNDEZ PERIS,
IRENE CARRILLO MURCIA
MERCEDES GUILABERT
Universidad Miguel Hernández

& EQUIPO DE INVESTIGACIÓN LENA[1]
Universidad Miguel Hernández, FISABIO y Universidad de Sevilla

1. INTRODUCCIÓN

El presente capítulo ofrece una visión detallada del proyecto *LENA - Investigación sobre sesgos de género en la generación de imágenes por Inteligencia Artificial*. Este proyecto, una colaboración entre la Universidad Miguel Hernández de Elche y el Instituto de las Mujeres del Ministerio de Igualdad de España, tiene como objetivo principal identificar y mitigar los sesgos de género presentes en los sistemas de inteligencia artificial generativa (IAG) que crean imágenes. La investigación se centra en comprender cómo y por qué estos sesgos se manifiestan en los modelos de IAG, profundizando en las estructuras subyacentes y los conjuntos de datos que alimentan dichos sistemas. Estos sesgos, que reflejan y amplifican las desigualdades presentes en la sociedad, no solo afectan la representación de los géneros en el ámbito digital, sino que

[1] El equipo de investigación LENAE está formado por Rosario Carmona Paredes, Ángela Coves Soler, Miguel Onofre Martínez Rach, José Joaquín Mira Solves, Victoria Soto Sanz (Universidad Miguel Hernández); Eva Gil Hernández, Daniel García Torres (FISABIO); Almudena Arroyo Rodríguez y María Calderón Fernández (Universidad de Sevilla)

también refuerzan estereotipos perjudiciales que pueden tener un impacto real en la vida cotidiana de las personas.

La relevancia de este proyecto radica en su enfoque innovador para identificar las barreras tecnológicas y socioculturales que perpetúan estas distorsiones de género. Además de revelar la presencia de sesgos en diversas plataformas de IA generativa, el proyecto LENA busca desarrollar soluciones concretas para su mitigación, explorando desde la intervención en los algoritmos hasta la modificación de los conjuntos de datos utilizados para el entrenamiento de estos sistemas. Con la creciente adopción de estas tecnologías en múltiples sectores, desde el marketing hasta la educación, se vuelve urgente garantizar que las imágenes generadas por IA sean inclusivas y respeten la diversidad de género. Así, el proyecto LENA no solo se inscribe en la corriente actual del feminismo digital, sino que también propone un marco de actuación para lograr una mayor equidad y justicia en el diseño y uso de tecnologías avanzadas.

Este capítulo examina, por tanto, los objetivos, las metodologías y los hallazgos preliminares del proyecto, contextualizando su importancia en el panorama actual de la tecnología y el feminismo, y demostrando cómo la lucha por la igualdad de género se traslada al ámbito de la inteligencia artificial. A medida que las tecnologías de IA se integran más profundamente en nuestras vidas, es fundamental que estas herramientas se desarrollen de manera responsable y que reflejen los valores de equidad, justicia y diversidad.

1.1. Marco del proyecto

La tecnología de inteligencia artificial ha avanzado significativamente, permitiendo la creación de modelos generativos que producen imágenes realistas a partir de descripciones textuales. Estos avances han sido posibles gracias al desarrollo continuo de algoritmos de aprendizaje profundo o *deep learning* (LeCun 2015), que han transformado la tarea de convertir texto en imágenes en una de las aplicaciones más sorprendentes y útiles en el campo de la visión por computadora (Li 2019, Mansinov 2016, Ramesh 2021, Reed 2016, Zhang 2017). Estos modelos no solo permiten la generación de imágenes a partir de descripciones

sencillas, sino que también pueden replicar estilos artísticos, crear imágenes fotorealistas y componer escenas complejas que parecen sacadas de la realidad.

Un ejemplo claro de esta capacidad se puede ver en la figura 1, que muestra una imagen generada por una IA, utilizando el modelo DALL-E de ChatGPT 4.0. En este caso, el prompt utilizado fue: *"Crea una imagen de un gato sentado en una silla verde en la playa y comiendo una hamburguesa gigante."*. El resultado es una representación visualmente interesante que refleja la capacidad de los modelos de IA para generar imágenes de alta calidad a partir de simples instrucciones textuales.

FIGURA 1. *Ejemplo de generación de imágenes con IA: mediante DALL-E 2 obtenemos una imagen artística de un gato en la playa con una hamburguesa gigante.*

Fuente: elaboración propia con DALL-E 2

Sin embargo, a pesar de su capacidad técnica, estos modelos no están exentos de limitaciones, especialmente en lo que respecta a los sesgos

inherentes presentes en los datos de entrenamiento. Los sesgos de género son un ejemplo prominente, que se reflejan a menudo en las imágenes generadas por estos sistemas. Por ejemplo, en la figura 2, que fue generada a partir del prompt: *"Crea una imagen de una profesora de matemáticas explicando qué es ChatGPT, Gemini, Copilot y Grok, delante de una pizarra. La profesora viste pantalones*". Aunque el prompt simplemente describe a una profesora, la imagen generada refleja estereotipos de género, mostrando a la profesora con curvas pronunciadas y tacones de aguja, reforzando representaciones sesgadas sobre el aspecto de las mujeres. Este tipo de sesgos en la generación de imágenes puede tener efectos perjudiciales, perpetuando estereotipos visuales en contextos donde la neutralidad debería prevalecer.

FIGURA 2. *Ejemplo de generación de imágenes con IA: mediante DALL-E 2 obtenemos una imagen de una profesora delante de una pizarra en un aula.*

Fuente: elaboración propia con DALL-E 2

En contraste, la figura 3 muestra la imagen generada a partir del prompt: "Crea una imagen de dos profesionales sanitarios, un hombre y una

mujer, charlando en una habitación de hospital al lado de la cama de un paciente.” En esta imagen, no se observa la presencia de estereotipos de género marcados, ya que no se generaron los estereotipos clásicos de mujer como enfermera con cofia y hombre como médico.

FIGURA 3. *Ejemplo de generación de imágenes con IA: mediante DALL-E 2 obtenemos una imagen de dos profesionales sanitarios en un contexto de hospital.*

Fuente: elaboración propia con DALL-E 2

Los ejemplos anteriores muestran de manera evidente que, aunque las plataformas de IA generativa tienen un enorme potencial, también existen riesgos relacionados con la reproducción de estereotipos de género. Sin embargo, estas plataformas también incluyen mecanismos, como filtros y ajustes algorítmicos, que permiten mitigar algunos de estos sesgos. Aun así, es evidente que queda mucho por hacer para perfeccionar estos sistemas y garantizar que sean inclusivos y no discriminatorios.

Estudios previos han demostrado que los sesgos en los datos de entrenamiento pueden llevar a la generación de imágenes que refuerzan estereotipos de género.

1.2. Estudios previos analizados

El trabajo de (García-Ull & Melero-Lázaro 2023) explora el sesgo de género en el lugar de trabajo en imágenes generadas por DALL-E, utilizando un muestreo probabilístico estratificado basado en 37 profesiones, se generaron 666 imágenes evaluadas en una escala Likert de 3 puntos. Se encontró que el 21.6% de las imágenes mostraban estereotipos completos de mujeres y el 37.8% de hombres. Los resultados indican que la IA no solo replica los estereotipos de género existentes, sino que los refuerza, con un 59.4% de imágenes mostrando estereotipos fuertes, comparado con el 35% en estudios con humanos.

En un trabajo similar, (Gorska & Jemielniak 2023) exploran el sesgo de género en imágenes generadas por IAG de profesionales, enfocándose en la representación visual de hombres y mujeres en derecho, medicina, ingeniería e investigación científica. Con una muestra de 99 imágenes de nueve generadores de texto a imagen y una encuesta a 120 personas, se encontró un sesgo significativo: los hombres representados en el 76% de las imágenes y las mujeres en solo el 8%.

Otro trabajo significativo en este ámbito es el publicado en Nature por (Kalluri 2024) en el que se exploran los sesgos de género y raza en imágenes generadas por IA. Este trabajo muestra que las herramientas de generación de imágenes como Stable Diffusion y DALL·E tienden a reproducir y amplificar estereotipos comunes, como asociar "África" con pobreza. Estos sesgos surgen de los datos sesgados en los que se entrenan estas IA. Aunque algunas empresas intentan contrarrestar estos sesgos, los resultados a menudo son insuficientes.

En el trabajo de (Sun et al. 2024) examinaron la prevalencia de dos tipos de sesgos de género en 15,300 imágenes de DALL·E, abarcando 153 ocupaciones. Los hallazgos revelan una subrepresentación de mujeres en campos dominados por hombres y una sobrerrepresentación en ocupaciones dominadas por mujeres, así como una tendencia a mostrar más mujeres sonrientes y con cabezas inclinadas hacia abajo.

Estos estudios previos subrayan la necesidad y la relevancia de esta investigación que aquí proponemos: las actuales IAs generativas presentan sesgos de representación y estereotipos pronunciados. Por lo tanto, es crucial implementar intervenciones feministas para mitigar los posibles impactos de estas imágenes generadas por IA en la sociedad.

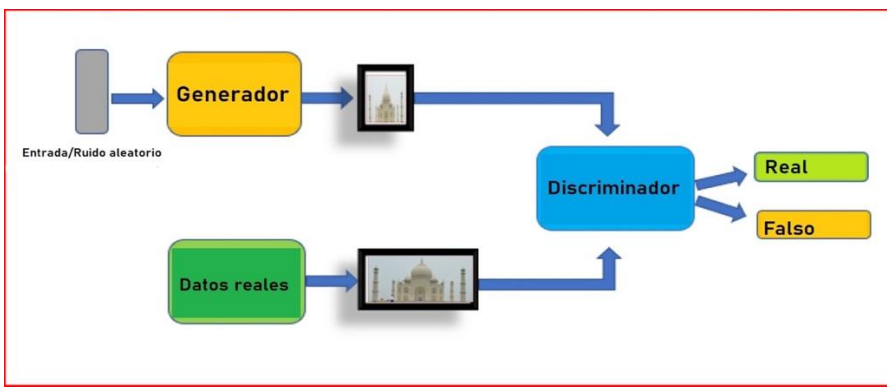
Estas intervenciones no solo deben buscar corregir los sesgos existentes, sino también promover una representación más equitativa y justa en las imágenes producidas por estas tecnologías avanzadas, contribuyendo a una sociedad más inclusiva y consciente de la diversidad.

1.3. Fundamentos Básicos de las Inteligencias Artificiales Generativa de Imágenes

Las IAG de imágenes, como DALL·E, se basan en técnicas avanzadas de aprendizaje profundo y redes neuronales generativas. Estas tecnologías incluyen redes generativas antagónicas (GANs) (Googfellow 2014) y modelos de difusión (Ho 2020), que aprenden a generar imágenes realistas a partir de descripciones textuales.

Las GANs consisten en dos redes: una generadora que crea imágenes y una discriminadora que evalúa su autenticidad. A través de un proceso de entrenamiento adversarial, ambas redes mejoran iterativamente. La figura 4 muestra un sencillo diagrama de flujo del funcionamiento de las GANs.

FIGURA 4. *Representación gráfica: Funcionamiento de una red GAN.*



Fuente: "Redes GAN: ¿Qué son? Características, funciones y ventajas"
https://redesinformaticas.org/red-gan/

Los modelos de difusión, por otro lado, emplean un enfoque probabilístico para generar imágenes detalladas a partir de ruido aleatorio, mejorando gradualmente la calidad de las imágenes (Ho 2020, Dhariwal 2021).

El éxito de estas IAGs depende de la calidad y diversidad de los datos de entrenamiento, así como de la capacidad del modelo para captar relaciones complejas entre texto e imagen. No obstante, los sesgos presentes en los datos de entrenamiento pueden influir en los resultados generados, subrayando la importancia de utilizar conjuntos de datos equilibrados y estrategias de mitigación de sesgos.

1.4. Listado de las principales Inteligencias Artificiales generadoras de imágenes

En este bloque, presentamos un listado de las principales inteligencias artificiales generativas de imágenes. Estas IAGs han revolucionado la forma en que se crean y manipulan imágenes, siendo capaces de generar una amplia variedad de estilos, desde representaciones fotorealistas hasta interpretaciones artísticas abstractas, incluyendo ilustraciones, logotipos y otros recursos visuales. La versatilidad y la calidad de las imágenes producidas han posicionado a estas plataformas como herramientas clave en sectores como el diseño gráfico, la publicidad, el arte y la educación.

A continuación, se destacan algunas de las IAGs más influyentes y populares en el ámbito general, utilizadas por miles de usuarios en todo el mundo:

DALL-E

Web oficial: https://openai.com/index/dall-e/

Descripción: Desarrollada por OpenAI y lanzada el 5 de enero de 2021, DALL-E se ha destacado como una de las principales IA generativas de imágenes. Capaz de generar imágenes desde descripciones textuales, ha llamado la atención por su capacidad para crear imágenes originales y detalladas basadas en prompts complejos y específicos.

Midjourney

Web oficial: https://www.midjourney.com/home

Descripción: Lanzada en beta abierta el 12 de julio de 2022, esta plataforma, creada por el equipo de David Holz, ha ganado rápidamente popularidad por su capacidad de producir imágenes de alta calidad en estilos artísticos y visuales variados. Midjourney ha sido particularmente aclamada en las comunidades artísticas por su capacidad para generar obras con un toque estilístico único.

Stable Diffusion

Web oficial: https://stability.ai/stable-image

Descripción: Creada por Runway y LMU Múnich, esta plataforma de código abierto fue lanzada en agosto de 2022. Stable Diffusion se distingue por su capacidad de generar imágenes de gran calidad de manera eficiente, lo que ha permitido que tanto profesionales como aficionados la utilicen en proyectos que van desde la creación de arte conceptual hasta la producción de contenidos comerciales.

Adobe Firefly

Web oficial: https://www.adobe.com/es/products/firefly.html

Descripción: Parte del ecosistema de Adobe Creative Cloud, Firefly fue lanzada en marzo de 2023 como una herramienta poderosa para la creación de imágenes asistida por IA. Ofrece una integración perfecta con las otras herramientas de Adobe, permitiendo a los diseñadores y artistas gráficos una mayor eficiencia en sus flujos de trabajo creativos, y se enfoca en garantizar imágenes de alta calidad, adaptadas a múltiples estilos y formatos.

Leonardo

Web oficial: https://leonardo.ai/

Descripción: Leonardo es una IA generativa conocida por su capacidad para producir imágenes con un nivel de detalle y calidad extraordinarios. Utilizada en el ámbito del arte digital y el diseño gráfico, Leonardo ha capturado la atención por su precisión en la creación de imágenes que abarcan desde retratos detallados hasta paisajes abstractos.

Canva

Web oficial: https://www.canva.com/ai-image-generator/

Descripción: Como una de las plataformas más populares de diseño gráfico, Canva ha incorporado un módulo de generación de imágenes con IA, denominado Magic Studio. Lanzado inicialmente en 2012 como una herramienta de diseño, Canva se ha adaptado rápidamente a las innovaciones en IA, ofreciendo a sus usuarios la capacidad de generar gráficos personalizados de manera sencilla y rápida.

A continuación, presentamos una selección de IAGs de imágenes diseñadas para usos más específicos, destacadas por su capacidad para abordar sectores o estilos particulares:

Comics Maker AI

Web oficial: https://www.comicsmaker.ai/

Descripción: Esta plataforma está especializada en la creación de imágenes para cómics. Ofrece herramientas que permiten a los artistas del cómic generar personajes, escenas y viñetas con estilos gráficos característicos de este género, reduciendo significativamente el tiempo de producción.

Scenario

Web oficial: https://www.scenario.com/

Descripción: Especializada en la creación de assets visuales para videojuegos, Scenario permite a los desarrolladores generar entornos, personajes y objetos específicos para sus proyectos. Su capacidad para generar contenido adaptado a la estética de los videojuegos la ha posicionado como una herramienta clave en la industria del entretenimiento interactivo.

HeyGen

Web oficial: https://www.heygen.com/

Descripción: HeyGen es una IA que se especializa en la creación de videos utilizando avatares predefinidos. Esta herramienta es utilizada principalmente en la producción de contenidos audiovisuales personalizados, permitiendo a los usuarios generar vídeos promocionales, educativos o de entretenimiento sin la necesidad de grandes equipos de producción.

2. OBJETIVOS

El propósito principal del proyecto LENA es investigar y abordar los sesgos de género que emergen en los sistemas de inteligencia artificial dedicados a la generación de imágenes. El proyecto busca comprender de manera profunda cómo y por qué estos sesgos se manifiestan en los modelos generativos utilizados, y tiene como finalidad el desarrollo de soluciones que permitan identificar y cuantificar de forma precisa la presencia de estos sesgos en diversas plataformas de IA. La investigación no solo se centra en evidenciar estas disparidades, sino también en proponer estrategias concretas para mitigar sus efectos.

Dentro de los objetivos generales, se plantea la creación de una metodología robusta para identificar y evaluar los sesgos de género en las imágenes generadas por IA. Esta metodología servirá como una herramienta clave para mejorar la equidad de género en el ámbito tecnológico, sensibilizando tanto a la comunidad científica como a los desarrolladores sobre la imperiosa necesidad de evitar y corregir estos sesgos. La meta última es promover el desarrollo de tecnologías más inclusivas y justas.

En cuanto a los objetivos específicos, se aspira a realizar una revisión exhaustiva de la literatura existente en torno a los sesgos de género en la generación de imágenes por inteligencia artificial. Esta revisión proporcionará el marco teórico necesario para diseñar y llevar a cabo experimentos prácticos que permitan evaluar los sesgos presentes en los modelos generativos actuales. A partir de estos experimentos, se desarrollarán y probarán técnicas destinadas a reducir y, en la medida de lo posible, eliminar los sesgos identificados. Finalmente, los resultados obtenidos serán difundidos ampliamente a través de publicaciones científicas y presentaciones en congresos, con el fin de compartir el conocimiento generado y fomentar una mayor sensibilización sobre la importancia de erradicar el sesgo de género en las tecnologías de inteligencia artificial.

3. METODOLOGÍA

La metodología del proyecto LENA se basa en una revisión exhaustiva de la literatura existente sobre los sesgos de género en la generación de imágenes por inteligencia artificial. Posteriormente, se diseñan y validan diversos *prompts* o instrucciones que representan situaciones y personas diversas. Estos prompts se utilizan en generadores de imágenes como DALL-E y Midjourney, y los resultados obtenidos se ajustan mediante pruebas piloto. A continuación, se aplican técnicas de análisis de imágenes para detectar la presencia de sesgos de género en un conjunto de datos representativo. También se realizarán evaluaciones ciegas de las imágenes generadas por IA, comparándolas con imágenes reales, utilizando un grupo diverso de evaluadores humanos.

3.1. Preguntas de investigación

1. ¿Es posible evaluar y mitigar el sesgo de género en las imágenes generadas por IA mediante una metodología adecuada?

2. ¿Son suficientes las técnicas actuales de supervisión para abordar eficazmente los sesgos de género en imágenes generadas por IA?

3.2. Hipótesis de trabajo

Las imágenes generadas por IA a partir de descripciones sobre personas incorporarán sesgos y estereotipos de género. Asimismo, se espera que las técnicas de análisis de imágenes permitan abordar y reducir eficazmente estos sesgos.

3.3. Fases de la metodología

A continuación de detallas las tareas a realizar en este estudio

3.3.1. Fase 1: Diseño y aplicación de prompts

En la primera fase, se revisa la literatura existente sobre los sesgos de género en imágenes generadas por IA y se diseñan prompts para evaluar estos sesgos en los modelos generativos. A través de pruebas piloto, se ajustan los prompts para optimizar los resultados.

3.3.2. Fase 2: Análisis del contenido generado

Durante esta fase, se analizarán las imágenes generadas por los modelos de IA utilizando técnicas de análisis de imágenes para identificar los sesgos de género. Se recopilará un conjunto de datos representativo y se implementarán algoritmos para asegurar la precisión de los resultados.

3.3.3. Fase 3: Desarrollo de una metodología de evaluación

A partir del análisis de las fases anteriores, se diseñará un protocolo para evaluar el sesgo de género en imágenes generadas por IA. Este protocolo será documentado de manera clara para su futura implementación en estudios adicionales.

3.3.4. Fase 4: Evaluación ciega

Se llevará a cabo una evaluación ciega de imágenes generadas por IA, comparándolas con imágenes reales. Un grupo diverso de evaluadores analizará las imágenes, lo que permitirá identificar los sesgos presentes y diseñar métricas para evaluar la efectividad de la metodología propuesta.

3.3.5. Fase 5: Difusión de resultados

En la última fase, se difundirán los resultados del proyecto a través de publicaciones científicas y presentaciones en congresos, asegurando su transferencia a la sociedad. El objetivo es sensibilizar a la comunidad tecnológica y al público sobre la importancia de eliminar los sesgos de género en las tecnologías de inteligencia artificial.

4. RESULTADOS PRELIMINARES

En el momento de la redacción de este trabajo, el proyecto LENA se encuentra en sus primeras fases de desarrollo (Fases 1 y 2), lo que implica que todavía estamos recopilando y analizando gran parte de los datos. Sin embargo, los primeros hallazgos, junto con la revisión exhaustiva de la bibliografía previa, ya nos ofrecen indicios claros sobre la presencia de sesgos de género en las imágenes generadas por inteligencia artificial a partir de descripciones textuales.

Uno de los aspectos más relevantes que hemos identificado hasta ahora en el análisis bibliográfico de los trabajos ya publicados es que los sistemas de IA, al transformar texto en imagen, tienden a reproducir estereotipos de género profundamente arraigados en la cultura. Por ejemplo, se suelen asociar ciertas profesiones, roles sociales o comportamientos con un género en particular. Así, trabajos como el de "enfermera" o "profesor de guardería" son representados mayoritariamente por mujeres, mientras que profesiones como "ingeniero" o "cirujano" se visualizan en su mayoría con hombres **(García-Ull & Melero-Lázaro 2023, Sun et al. 2024)**. Esta tendencia refleja cómo los sesgos presentes en los datos de entrenamiento, que suelen estar basados en estructuras sociales y culturales tradicionales, influyen directamente en los resultados generados.

FIGURA 5. *Ejemplo de generación de imágenes con IA: mediante DALL-E 2: "Crea 4 imágenes, en cada una de ellas que aparezca un solo personaje, hombre o mujer, y que esté desarrollando una actividad profesional"*

Fuente: elaboración propia con DALL-E 2

En cuanto a nuestras propias pruebas realizadas específicamente con DALL-E 2 (durante Junio-Septiembre 2024), los resultados son alentadores en relación con la neutralidad de género en profesiones tradicionalmente asociadas a un género específico. Tras la implementación de ciertos *prompts* cuidadosamente diseñados, hemos observado que DALL-E 2 es capaz de generar imágenes que muestran a hombres y mujeres de manera equitativa en profesiones como la ingeniería, la enfermería o el liderazgo empresarial, rompiendo con los estereotipos culturales preexistentes. Este avance sugiere que, aunque los sesgos siguen presentes, es posible reprogramar o ajustar los modelos generativos para evitar la reproducción de estos estereotipos, lo que es un paso importante hacia la creación de IA más justas e inclusivas.

La figura 5 presenta algunos ejemplos de estas imágenes procedentes de DALL- E2, donde se puede observar una neutralidad de género en la representación de profesiones que históricamente han sido dominadas por uno u otro género. Estos resultados preliminares son un indicio prometedor de que, con las metodologías y ajustes adecuados, podemos influir positivamente en cómo la IA representa a las personas, promoviendo una visión más equitativa y menos estereotipada.

5. DISCUSIÓN

5.1. ¿Por qué el acrónimo LENA?

El equipo de investigación ha decidido utilizar el acrónimo LENA para esta propuesta como un homenaje a Lena Söderberg (The Lena Story 2024), cuya imagen se ha convertido en un símbolo icónico dentro de la historia del procesamiento digital de imágenes. La fotografía de Lena, extraída de una revista Playboy de noviembre de 1972, ha sido utilizada durante décadas como referencia en el ámbito de la visión por computadora, siendo empleada para probar y comparar algoritmos de procesamiento de imágenes debido a su composición visualmente compleja y adecuada para los desafíos técnicos de la época. Sin embargo, el uso prolongado de esta imagen también ha sido objeto de controversia, ya que resalta las limitaciones y los sesgos de género inherentes a la tecnología y su desarrollo.

FIGURA 6. *Lena Söderberg, una modelo sueca de Playboy, es el rostro de la imagen de prueba más utilizada en el mundo. (Un ejemplo de filtros de procesado sobre Lena)*

Fuente: Elaboración propia con Matlab

El impacto histórico de la imagen de Lena no puede ignorarse, pero su uso también pone de relieve las tensiones entre el avance tecnológico y los valores sociales. En este sentido, nuestro estudio busca darle a Lena una "revancha" simbólica, convirtiéndola en un emblema de lucha contra los sesgos en la tecnología. No se trata solo de ajustar los algoritmos; es también un esfuerzo por replantear el enfoque con el que se desarrollan las herramientas tecnológicas, promoviendo una representación más justa y equitativa de todas las personas, independientemente de su género, raza o contexto social. Lena, en este nuevo contexto, representa la posibilidad de transformación: de ser un ejemplo de sesgo de género a convertirse en un icono de igualdad y diversidad en el ámbito digital.

5.2. Limitaciones

A pesar del alcance y la importancia de esta investigación, somos conscientes de las limitaciones inherentes a nuestro enfoque. En primer lugar, la variabilidad en los *prompts* utilizados para generar las imágenes puede influir considerablemente en los resultados obtenidos. Diferentes formulaciones de los prompts, por sutiles que sean, pueden conducir a imágenes muy distintas, lo que dificulta la generalización de los hallazgos. Por otro lado, la evaluación subjetiva de las imágenes generadas

introduce un factor de variabilidad adicional, ya que las percepciones individuales de los evaluadores pueden influir en los resultados.

Además, los modelos de IAG están en constante evolución. A medida que las versiones de los algoritmos se actualizan, los sesgos que identificamos en este estudio podrían cambiar o ser mitigados en futuras iteraciones. Por lo tanto, los resultados obtenidos deben interpretarse en el contexto temporal en el que se realizaron los experimentos, sabiendo que los avances en IAG podrían alterar las conclusiones a largo plazo.

Otra limitación importante es que los estereotipos de género pueden variar significativamente entre diferentes contextos culturales. Lo que se considera un sesgo de género en una cultura podría no percibirse de la misma manera en otra. Esta diversidad cultural implica que los resultados de este estudio pueden no ser universalmente aplicables, lo que abre la puerta a investigaciones más profundas y específicas según el contexto cultural.

A pesar de estas limitaciones, creemos que nuestro trabajo ofrece una contribución relevante al campo de la inteligencia artificial, proporcionando una base sólida para futuras investigaciones sobre los sesgos de género en imágenes generadas por IAG. La identificación de estas limitaciones también abre nuevas líneas de investigación que pueden ayudar a superar los desafíos actuales y avanzar hacia un futuro más inclusivo y diverso en el ámbito tecnológico.

6. CONCLUSIONES

El proyecto LENA pretende poner de manifiesto la urgente necesidad de abordar los sesgos de género presentes en los sistemas de inteligencia artificial generativa, particularmente en la creación de imágenes. A través de la revisión exhaustiva de la literatura, el diseño de *prompts* específicos y la implementación de metodologías de análisis de imágenes, queremos evidenciar que los sesgos de género pueden persistir en las imágenes generadas artificialmente.

Este estudio también quiere poner en relieve las limitaciones a las que nos enfrentamos. Los sesgos culturales, la subjetividad en la evaluación

de imágenes y la constante evolución de los modelos de IA son aspectos que limitan la universalidad de nuestros hallazgos. Sin embargo, estos desafíos también representan oportunidades para futuras investigaciones que puedan refinar y expandir las metodologías que hemos propuesto.

En conclusión, el proyecto LENA establece las bases para continuar explorando cómo la inteligencia artificial puede contribuir a una representación más justa e inclusiva. Sabemos que el camino hacia una IA verdaderamente equitativa es largo y complejo, pero estamos convencidos de que la combinación de un enfoque interdisciplinar, una sensibilización crítica y el compromiso ético en el desarrollo de estas tecnologías puede marcar una diferencia significativa en el futuro de la inteligencia artificial y su impacto en la sociedad.

7. AGRADECIMIENTOS/APOYOS

Quisiéramos expresar nuestro más sincero agradecimiento a todas las personas e instituciones que han contribuido de manera significativa a la realización de este proyecto. En primer lugar, agradecemos a la Universidad Miguel Hernández por su apoyo incondicional y por proporcionar los recursos necesarios para llevar a cabo esta investigación. Asimismo, extendemos nuestro profundo agradecimiento al Instituto de las Mujeres del Ministerio de Igualdad por concedernos subvenciones para la realización de este estudio dentro de su convocatoria de Investigaciones Feministas de 2024. Su apoyo es fundamental para el desarrollo y éxito de esta investigación.

8. REFERENCIAS

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *arXiv preprint arXiv:2105.05233*.

García-Ull, F.J., & Melero-Lázaro, M. (2023). Gender stereotypes in AI-generated images. Profesional De La información, 32(5). https://doi.org/10.3145/epi.2023.sep.05

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Gorska, A. M., & Jemielniak, D. (2023). The invisible women: uncovering gender bias in AI-generated images of professionals. Feminist Media Studies, 23(8), 4370–4375. https://doi.org/10.1080/14680777.2023.2263659

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33, 6840-6851.

Kalluri, P. R. (2024). AI image generators often give racist and sexist results: Can they be fixed? *Nature*, 627, 722-725. https://doi.org/10.1038/d41586-024-00599-8

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

Li, B., Qi, X., Lukasiewicz, T., & Torr, P. (2019). Controllable text-to-image generation. *Advances in Neural Information Processing Systems, 32*.

Mansimov, E., Parisotto, E., L. Ba, J., & Salakhutdinov, R. (2016). Generating images from captions with attention. *ICLR*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In *ICML*.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In *International Conference on Machine Learning* (pp. 1060-1069). PMLR.

Sun, L., Wei, M., Sun, Y., Suh, Y. J., Shen, L., & Yang, S. (2024). Smiling women pitching down: Auditing representational and presentational gender biases in image-generative AI. *Journal of Computer-Mediated Communication, 29*(1).

The Lenna Story - www.lenna.org (2024)

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5907-5915).