EDICIONES COMPLUTENSE

Investigaciones Feministas
ISSN-e: 2171-6080

MONOGRÁFICO

Posibilidades de la Inteligencia Artificial (IA) para la prevención de la violencia de género

Iolanda Tortaja
Universitat Rovira i Virgili (España)
Cilia Willem
Universitat Rovira i Virgili (España)
Rosa María Gil Iranzo
Universidad de Lleida (España)

https://dx.doi.org/10.5209/infe.100602 Recibido: Enero 2025 • Evaluado: Marzo 2025 • Aceptado: Mayo 2025

ES **Resumen: Introducción y objetivos:** La aparición masiva de herramientas de inteligencia artificial (IA) conlleva un desafío importante para las investigaciones feministas: sus sesgos de género en detrimento de las mujeres. Este estudio explora las posibilidades de la IA como herramienta en la detección de la violencia de género en las películas y las series a través del análisis fílmico, usando el test de Bechdel-Wallace. **Metodología**: Para ello se diseñó un cuestionario de autoevaluación basado en el test de Bechdel-Wallace y en otros modelos que analizan la representación de las mujeres en los productos audiovisuales. También se utilizó la IA generativa Copilot (v.2024) para diseñar un cuestionario, y sus resultados fueron comparados con los de un grupo de 29 estudiantes universitarios. Copilot (v.2024) elaboró un cuestionario que está en línea con los principios de prevención y sensibilización establecidos por el equipo investigador, una muestra de que la IA, cuando es guiada y supervisada, puede generar herramientas útiles para el análisis crítico de la representación de género. Los estudiantes aplicaron el cuestionario a productos culturales de su elección, tomando una distancia crítica que favoreció la identificación de violencias machistas y la reflexión sobre la representación de género en los medios. **Resultados**: se compararon las respuestas del grupo piloto con las generadas por Gemini 1.5 y se encontró un 78% de coincidencia entre ambas. No obstante, en un 13% de los casos, las respuestas humanas mostraron un análisis más profundo y matizado que las de la IA, que tendía a formular respuestas estandarizadas sin captar plenamente la complejidad de ciertas dinámicas de poder. **Implicaciones prácticas:** El estudio sugiere que la IA generativa puede ser una herramienta de apoyo para la educación y la prevención de la violencia de género siempre que su uso esté supervisado y complementado con análisis humanos críticos y reflexivos, basados en la teoría y las investigaciones feministas.
Palabras clave: inteligencia artificial, test de Bechdel-Wallace, violencia de género, sesgos de género, análisis fílmico, feminist media studies, prevención

ENG Possibilities of Artificial Intelligence (AI) for the prevention of gender-based violence

Abstract: Introduction and aims: The massive emergence of artificial intelligence (AI) tools presents a significant challenge for feminist research: their gender biases to the detriment of women. This study explores the potential of AI as a tool for detecting gender-based violence in films and series through film analysis by means of the Bechdel-Wallace test. **Methodology**: A self-assessment questionnaire was designed, based on the Bechdel-Wallace test and other models that analyze the representation of women in audiovisual products. The generative AI Copilot (v.2024) was used to design a questionnaire, and its results were compared with those of a group of 29 university students. The questionnaire developed by Copilot (v.2024) aligned with the prevention and awareness principles established by the research team, demonstrating that AI, when guided and supervised, can generate useful tools for the critical analysis of gender representation. The students applied the questionnaire to cultural products of their choice, maintaining a critical distance that facilitated the identification of gender-based violence and reflection on gender representation in the media. **Results**: The responses from the pilot group were compared with those generated by Gemini 1.5, yielding a 78% coincidence between both. However, in 13% of the cases, human responses showed a deeper and more nuanced analysis than those of the AI, which tended to provide standardized answers without fully capturing the complexity of certain power dynamics. **Implications**: The study suggests that generative AI

can be a supportive tool in educational contexts and the prevention of gender-based violence, provided its use is supervised and complemented with critical and reflective human analysis, grounded in feminist theory and research.
Keywords: artificial intelligence, Bechdel-Wallace test, gender-based violence, gender bias, film analysis, feminist media studies, prevention

Cómo citar: Tortaja, I.; Willem, C.; Gil Iranzo, R. M. (2025). Posibilidades de la Inteligencia Artificial (IA) para la prevención de la violencia de género. *Investigaciones Feministas, 16*(1), 61-70. https://dx.doi.org/10.5209/infe.100602

1. Introducción

En los últimos años, el interés por la inteligencia artificial (IA) ha crecido de manera exponencial. En el momento de escribir estas líneas, la empresa china DeepSeek acaba de lanzar su primera aplicación de bot conversacional gratuita. En solo 17 días, DeepSeek-R1 superó a ChatGPT como la app más descargada en la App Store de iOS en Estados Unidos. La vertiginosa evolución de esta tecnología viene impulsada por un fuerte interés económico y político. De hecho, la IA no solo promete revolucionar la industria tecnológica, sino que también plantea importantes desafíos para la sociedad en su conjunto. Livingston (2023) y Nowotny (2023) destacaron precisamente esa dualidad de la IA: por un lado, ofrece oportunidades sin precedentes para la innovación y el progreso; por otro, plantea riesgos que deben ser gestionados cuidadosamente para evitar consecuencias negativas.

En primer lugar, el desarrollo y la implementación de la IA requieren de una gran cantidad de recursos, tanto en términos energéticos como de materiales. Crawford (2021) ya demostró que la infraestructura necesaria para sostener los sistemas de IA es considerable, lo que subraya la necesidad de abordar cuestiones de sostenibilidad y eficiencia en el uso de estos recursos. Pero uno de los aspectos más críticos de la IA, y el desafío más importante para las investigaciones feministas, es su tendencia a propiciar los consabidos sesgos de género en detrimento de las mujeres. Estos sesgos surgen debido a que la IA –los modelos de lenguaje, el aprendizaje automático (AA) y la IA generativa– se basa en datos que, a menudo, son ya inherentemente tendenciosos en origen. La máquina nos devuelve las palabras, los cálculos y las imágenes que le hemos proporcionado a partir de nuestras publicaciones y contenidos previos en internet. Como resultado, la IA perpetúa y amplifica las desigualdades de género existentes, por ejemplo, en ámbitos como el empleo, la educación y el acceso a servicios financieros, pero también en el plano simbólico, en la producción cultural y la creación de significados.

En este artículo, sin embargo, también queremos dejar claro que la IA tiene a su vez un elevado potencial como herramienta hipereficiente en la lucha contra los estereotipos y la violencia de género. Su capacidad para analizar grandes volúmenes de datos y detectar patrones ocultos ofrece, por ejemplo, una oportunidad para abordar estos problemas de manera más efectiva y temprana. Aunque la IA presenta desafíos significativos por su tendencia a perpetuar la generización, también ofrece herramientas innovadoras para combatir la violencia de género y promover la igualdad. Por eso es crucial que el desarrollo y la implementación de la IA se realicen con una perspectiva de género a fin de maximizar sus beneficios y minimizar sus riesgos.

Este artículo presenta parte de los resultados de un experimento realizado en 2024 con herramientas de IA para el análisis fílmico de productos culturales actuales desde un punto de vista de violencia de género. Concretamente, en este experimento pusimos a prueba la IA con una versión del test de Bechdel centrada en la detección de las violencias machistas en la ficción, y comparamos sus respuestas con las de participantes humanos. Para descartar efectos intrínsecos a un modelo de lenguaje a gran escala (LLM) concreto, hemos utilizado dos herramientas de IA diferentes: Copilot (v.2024) y Gemini (v1.5). Escogimos así los bots de dos de las mayores compañías punteras en IA como son Microsoft y Google, pero no ChatGPT (de OpenAI), dado que los productos que se utilizan en el ámbito académico para realizar documentos pertenecen mayoritariamente a Microsoft (OneDrive) y Google (Drive).

2. Los riesgos del sesgo de género en la IA

La inteligencia artificial, el aprendizaje automático o la robótica han dejado de ser conceptos abstractos o excesivamente técnicos para convertirse en realidades tangibles que dan forma de manera muy concreta a la sociedad, la economía y la cultura actuales, y en definitiva a nuestras vidas, irremediablemente interconectadas (Schwab 2016, Castells 2010). Sin embargo, estas tecnologías 'inteligentes' dan forma de manera diferente a nuestra realidad cotidiana dependiendo de si somos mujeres u hombres, pobres o ricos, del norte global o del sur global. En este sentido, un elemento particularmente perjudicial es el de los sesgos de

género, que afectan a más de la mitad de la población mundial. La construcción social del género, generada a través de la representación simbólica (Hall, 1973; Morley, 1996; Kaplan, 1998), juega un papel fundamental en la formación de las interacciones y las definiciones de la personalidad 'generizada' (Butler, 1990) y es moldeada por fuerzas sociopolíticas, económicas y tecnológicas (Connell, 2002).

El sesgo de género, según el Instituto Europeo de Igualdad de Género (2023), se define como "todas las acciones o los pensamientos basados en la percepción de que las mujeres no son iguales a los hombres en derechos ni en dignidad" (p. 2). Si consideramos el sesgo de género como una forma de violencia, esta incluye un amplio espectro de manifestaciones que pueden surgir en las interacciones humano-máquina (Deepanjali et al., 2024). Sirvan como ejemplos de estas la subrepresentación y marginación de las mujeres en campos relacionados con la tecnología (Ashcraft et al., 2016) o la reafirmación de estereotipos de género en los modelos de IA y aprendizaje automático (Crawford & Paglen, 2019).

Los modelos de lenguaje, como los utilizados en asistentes virtuales y chatbots, a menudo están entrenados con datos que reflejan un lenguaje androcéntrico, sexista y discriminatorio hacia las mujeres. Como reconoció en 2019 el fundador y CEO de Cogito, una empresa de IA que ofrece servicios a otras empresas: "(...) los sistemas de procesamiento de lenguaje natural (NLP), un ingrediente fundamental de los sistemas de IA comunes como Alexa de Amazon y Siri de Apple, entre otros, reproducen y amplifican sesgos de género presentes en los datos de entrenamiento" (Feast, 2019).

Esto puede dar como resultado respuestas y comportamientos que refuerzan estereotipos de género, lo que afecta negativamente a la percepción de las mujeres y al tratamiento que se les da en las interacciones tecnológicas, y por extensión en las interacciones interpersonales en la vida real. Como ejemplo, podemos recordar que se acusó a las IA de no ser inclusivas y cuando se intentó resolver ese problema, aparecieron inexactitudes históricas como que los modelos de IA generativa presentaban a soldados de las SS de raza negra o a vikingos asiáticos. En el último año, se ha puesto especial énfasis en evidenciar los sesgos de género (Fraile-Rojas, B. et al., 2025), aunque las IA no siempre nos devuelven imágenes realistas de mujeres, como se puede observar en la Figura 1.

Figura 1. Imagen estereotipada de mujeres "mayores" creada por la inteligencia artificial generativa Midjourney v.6.1.

Otra manifestación de esta tendenciosidad son los mensajes misóginos y antifeministas que proliferan en las redes sociales y los foros de internet, fruto de una cultura de la violación, y que se acaban filtrando en las herramientas de IA en un bucle exponencial. Los algoritmos de IA que analizan y gestionan contenido en plataformas digitales se dejan llevar por la prevalencia de estos mensajes, lo que propicia una moderación de contenidos sesgada, ineficaz o directamente inexistente. Esto no solo afecta (negativamente) a la experiencia de las usuarias en estas plataformas, sino que también contribuye a la normalización de la misoginia y la violencia contra las mujeres en el entorno digital.

Un ejemplo extremo son las herramientas y apps que permiten la creación de *porno deepfake* e imágenes *deepnude* que perjudican desproporcionadamente a las mujeres y las exponen a formas extremas de violencia y explotación digital. La generación automática de fotos y vídeos de cuerpos desnudos con rostros de personas existentes añade un nivel de violencia que es difícil de gestionar ante la dificultad de establecer la autoría de dichas imágenes y su rápida y descontrolada extensión, como ocurrió en el conocido caso de Almendralejo. Al ser un contenido pornográfico falso generado sin el consentimiento de las personas involucradas, puede tener graves consecuencias para la privacidad y la seguridad de las mujeres y las niñas (Smith & Rustagi, 2021). Además, la facilidad con la que se puede crear y distribuir este tipo de contenido ha

comportado un aumento de los casos de acoso y extorsión, y, por lo tanto, una exacerbación de la violencia de género en el ámbito digital.

También encontramos sesgos de género en la visibilización y la difusión de las contribuciones científicas. En muchos campos científicos, los modelos de referencia son predominantemente masculinos, lo que se traduce en una mayor citación de sus obras y en una invisibilización de las contribuciones de las mujeres a la ciencia. Este fenómeno, conocido como el "efecto Matilda", explica por qué las mujeres científicas a menudo reciben menos reconocimiento por sus trabajos que sus colegas masculinos (O'Connor & Liu, 2024). La IA, al basarse en estos datos sesgados, perpetúa esta desigualdad, pues prioriza la citación de obras masculinas en sus algoritmos de búsqueda y sus listados de referencias académicas. En un estudio experimental sobre la traducción automatizada, unos investigadores brasileños encontraron que las traducciones hechas por IA están fuertemente sesgadas hacia los valores predeterminados masculinos, especialmente en campos STEM, que típicamente se consideran inclinados hacia dicho género (Prates et al., 2020).

Las autoras O'Connor & Liu (2024) incluso señalan el peligro del uso de la IA en la toma de decisiones por parte de las administraciones públicas y en la gestión de las políticas en ámbitos como la sanidad y el sistema judicial. Hemos visto varios ejemplos ya de los 'algoritmos predictivos' para la automatización de decisiones en el sector público, como la denegación de ayudas a familias vulnerables en Reino Unido en 2020 (Human Rights Watch, 2020), o la predicción del crimen como base de las decisiones policiales en los Estados Unidos, los Países Bajos y también en el Reino Unido. En este sentido, O'Connor & Liu alertan de que, aunque la IA en sí misma puede ser vista como una tecnología neutral y objetiva, esta adquiere nuevos significados e implicaciones a través de su uso humano en contextos específicos (2024: 2046). No es difícil imaginarse cómo un sesgo de género en la gestión de casos judiciales o de litigios basada en herramientas de IA puede ser perjudicial, peligroso o directamente letal para las mujeres, especialmente las que pertenecen a colectivos vulnerables.

Todos estos ejemplos subrayan la urgencia para abordar y mitigar los sesgos de género en la IA a fin de garantizar que estas tecnologías promuevan la igualdad y no perpetúen las situaciones de violencia. La investigación y el desarrollo de IA deben centrarse en la creación de sistemas más inclusivos y equitativos que reflejen una diversidad de experiencias y perspectivas.

3. Oportunidades y brechas para utilizar la IA en la prevención de la violencia de género

Igual que la IA contiene los sesgos de género que acabamos de exponer y que son consecuencia de lo que le hemos transmitido nosotros, los seres humanos, también puede desempeñar un papel crucial en la identificación y la mitigación de estereotipos de género. Del mismo modo que los algoritmos predictivos pueden automatizar decisiones policiales con consecuencias desastrosas, las IA también pueden ser utilizadas para prevenir casos de violencia de género mediante la identificación temprana de ciertos comportamientos y la intervención antes de que se produzca una situación de violencia. Un proyecto de innovación reciente de la Universitat Oberta de Catalunya propone la arquitectura de un sistema que ayude a detectar posibles situaciones de violencia de género construyendo perfiles detallados de agresores y víctimas con base en casos anteriores (Plo-Moreno, 2023).

Este sistema podría permitir a las fuerzas de seguridad del Estado y a los jueces determinar zonas geográficas y momentos óptimos para llevar a cabo acciones concretas a nivel preventivo. También podría ayudar a otras instituciones y organismos públicos a realizar acciones de formación y sensibilización en materia de violencia de género. De la misma manera, las plataformas de IA pueden ofrecer apoyo a las víctimas de esa violencia mediante chatbots y asistentes virtuales que proporcionen información o recursos en tiempo real. Además, pueden ayudar a los servicios públicos a identificar y responder a situaciones de violencia de manera más rápida y eficiente (Guthridge et al., 2022).

En el contexto educacional, Barrera Yañez et al. (2023) sugieren que los juegos serios (*serious games*) diseñados con IA pueden ayudar a abordar los estereotipos de género en entornos educativos. Si se hacen con conocimiento de causa, estos juegos permiten a los usuarios experimentar diferentes escenarios y reflexionar sobre sus propios prejuicios y comportamientos, proporcionando una plataforma interactiva para la discusión y el aprendizaje colectivo sobre el género y la violencia. En otros escenarios, la IA se podría usar para diseñar y desarrollar programas formativos o planes docentes que desafíen los estereotipos de género y promuevan la igualdad. Por ejemplo, la incorporación de IA en la enseñanza de lenguaje *queer* ha mostrado ser efectiva para prevenir la violencia de género al fomentar un ambiente inclusivo y respetuoso (Palacios & Huertas, 2024).

Conviene destacar, por último, el potencial de las herramientas IA para propiciar un cambio cultural. Como hemos visto en los párrafos anteriores, la IA no deja de ser un reflejo de nuestra cultura. En concreto, los productos culturales son uno de los ámbitos donde más se propagan los estereotipos de género. Autores provenientes de los Estudios Culturales británicos como Hall (1973) y Morley (1996) han demostrado con abundantes ejemplos que los relatos de los medios de comunicación, las películas y otros productos culturales pueden perpetuar estos estereotipos. Sin embargo, la IA tiene el potencial de desafiar y cambiar estas narrativas al crear contenido (audiovisual) que muestre una mayor diversidad y promueva la igualdad de género. En el vídeo "My Word"[1], creado especialmente para una exposición del CCCB[2], la artista visual Carme

[1] Véase https://www.carmepuche.com/my-word

[2] Véase https://www.cccb.org/es/exposiciones/ficha/ia-inteligencia-artificial/240941

Puche Moré muestra lo sesgada que está de partida la IA en la generación de imágenes, y 'entrena' el modelo para que tenga en cuenta la diversidad y reproduzca otro tipo de imágenes. Al final la IA termina presentando a una médica negra de edad avanzada, y rompe así con los estereotipos tradicionales al ofrecer una representación más inclusiva y, en definitiva, realista. "My Word" es un ejemplo canónico de propuesta artística crítica sobre la IA que contribuye a la generación de imágenes alternativas y, por ende, a un cambio en el relato colectivo.

En el experimento que llevamos a cabo exploramos estas posibilidades de la IA para el análisis crítico de productos culturales, en este caso, películas y series televisivas.

4. Metodología

El experimento se enmarca en una investigación más amplia llevada a cabo por el Ayuntamiento de Lleida y la Universitat de Lleida, en colaboración con investigadoras de la Universitat Rovira i Virgili. El proyecto explora los riesgos y las oportunidades de la IA en relación con la violencia de género para tratar de desarrollar políticas de prevención que aprovechen el potencial de las inteligencias artificiales. Para monitorizar algunos modelos de lenguaje y asistentes virtuales y poner a prueba la capacidad de la IA generativa como herramienta de prevención, se diseñó una herramienta en forma de cuestionario de autorrespuesta pensado para su difusión en redes sociales. Se tomó como punto de partida el conocido test de Bechdel-Wallace, que, desde su publicación en 1985, se ha popularizado mucho y se usa ya habitualmente para hacer análisis críticos sobre la presencia (o la ausencia) de mujeres en la pantalla. Centrado en tres preguntas sencillas, el test de Bechdel-Wallace permite identificar rápidamente qué películas o productos culturales cumplen un requisito mínimo en cuanto a la representación de las mujeres:

- ¿Hay por lo menos dos personajes femeninos con nombre?
- ¿Estas mujeres hablan entre sí?
- ¿Su conversación trata sobre algo que no sea un hombre?

Durante décadas, el cuestionario ha servido para reflexionar sobre la invisibilización de la mujer en la pantalla y denunciar los sesgos que se derivan de la mirada masculina y patriarcal imperante (y persistente aún) en los relatos cinematográficos (Mulvey, 1975; Kaplan, 1998).

Para completar nuestra propuesta, se revisaron otros cuestionarios inspirados en el test de Bechdel-Wallace, como son: *reverse Bechdel*, *Mako Mori*, *Sexy Lamp*, *Sphinx*, *Johanson analysis*, *Vito Russo*, *Deggans*, *Shukla*, *Riz*, *DuVernay*, *Nadia Latif and Leila Latif questions*, *Kent*, *Aila*, *Finkbeiner*. Además, para el proyecto, se tuvieron en cuenta algunas de las dimensiones del llamado "iceberg" de las violencias –las sutiles y las explícitas– contra las mujeres. Se incorporó así el papel del humor, el uso del lenguaje sexista, y la presencia de desprecio, control, amenazas y violencia directa/física. Finalmente, se utilizaron dos de las categorías desarrolladas por Donnerstein (1998) relativas a la representación atractiva del agresor y a la justificación o minimización de la violencia en el relato.

Después, se pidió al asistente virtual Copilot (v.2024) que diseñara un cuestionario (siguiendo los mismos pasos que habían seguido las investigadoras). Se le pidió también que recomendara un producto audiovisual ya estrenado en el mercado que contuviera algún tipo de representación alternativa de las violencias machistas: alguna película o serie que visibilizara un tratamiento complejo y no estereotipado del tema. En cuanto a la primera de esas tareas, el asistente desarrolló un cuestionario inesperadamente muy parecido al que diseñaron las investigadoras a partir de todos los test que ellas mismas habían explorado para incluir las dimensiones de la violencia descritas con anterioridad. Con ello, Copilot (v.2024) demostró que puede ser una herramienta que, con una conducción previa y una adecuada monitorización de todo el proceso, puede generar buenos cuestionarios.

A continuación, se le pidió a la IA que ampliara los ítems de dicho cuestionario incorporando otros recogidos de Donnerstein (1998): ausencia de violencia, motivos de la violencia, lenguaje y humor, y atractivo del agresor. La herramienta recomendó entonces una serie para aplicarle su propio cuestionario: *The Handmaid's Tale* (*El cuento de la criada,* en España). Además, a efectos de obtener respuestas más acordes con las de un público humano más joven, también se le pidió que probara a responder sus preguntas aplicadas a la serie *Euphoria*.

Una vez probado el cuestionario en la IA, este fue pasado a un grupo de 29 alumnos de la asignatura de "Gamificación y *Serious Games*", del segundo curso del grado de Diseño Digital y Tecnologías Creativas de la Universitat de Lleida. En el grupo –como en el grado en general– el número de chicos (H) y de chicas (D) estaba muy equilibrado. Además, una persona se definió como no binaria (NB) y también usamos el código "no procede" (NP) para aquellas que prefirieran no identificar su género. En cuanto al rango de edad, oscilaba entre los 19 y los 22 años.

Una vez que las personas hubieron contestado individualmente el cuestionario sobre la serie o la película que habían escogido, se pidió a la IA de Google, Gemini, que hiciera lo propio y contestara el cuestionario para las 28 series o películas escogidas por el alumnado participante (no fueron 29, porque una de las series de televisión se repetía [ver Tabla 1]) a fin de comparar sus respuestas con las humanas.

Tabla 1. Listado de series y películas analizadas en el experimento.

1. INTERSTELLAR
2. ROCKY HORROR PICTURE SHOW
3. FLEABAG
4. MUJERCITAS
5. TITANIC
6. CÓMO CONOCÍ A VUESTRA MADRE
7. PRISON BREAK
8. CÓMO CONOCÍ A VUESTRA MADRE
9. STAR WARS JEDI: THE FALLEN ORDER
10. SKYRIM
11. ONE PIECE
12. HORA DE AVENTURAS
13. THE OWL HOUSE
14. FIGHT CLUB
15. SKY ROJO
16. BEASTARS
17. MY LITTLE PONY
18. FRIENDS
19. CREPÚSCULO
20. UNDER THE SKIN 1
21. BREAKING BAD
22. BROOKLYN 99
23. LA CASA DRAGÓN
24. ATYPICAL
25. LAS CHICAS DEL CABLE
26. EVERYTHING EVERYWHERE ALL AT ONCE
27. VIS À VIS
28. GINNY Y GEORGIA
29. SEXO EN NUEVA YORK

5. Resultados

5.1. Coincidencia entre los guiones de los cuestionarios de las investigadoras y los de la IA

Uno de los resultados más relevantes de este pequeño experimento tiene que ver con la coincidencia de los cuestionarios elaborados por el equipo de investigadoras del proyecto y los de Copilot (v.2024). Se utilizaron diversos programas de inteligencia artificial para ver si eran capaces de generar preguntas significativas. Para intentar tener un panorama de inteligencias artificiales generativas más amplio en el estudio, decidimos utilizar Copilot (v. 2024) para la generación de las preguntas y Gemini 1.5 para la posterior comparación de las respuestas, dado que, de entrada, ambos bots ofrecían siempre rendimientos parecidos en sus respuestas.

En este caso, Copilot (v.2024) generó un texto fundamental desde el plano teórico porque elaboró un guión de preguntas no basado en los sesgos y estereotipos de género detectados en las investigaciones sobre IA (Crawford & Paglen, 2019; Deepanjali et al., 2024). Además, como el encargo inicial había sido muy concreto, se le pidió a la IA que extendiera el test de Bechdel-Wallace para que incorporara la perspectiva de género, de forma que las preguntas que desarrolló Copilot (v.2024) abordaran los mismos temas y estuvieran formuladas en los mismos términos que las del guion elaborado por el equipo investigador. Con todo ello, juzgamos pertinente usar para el piloto el test generado por Copilot (v.2024), dado que vimos que promueve un enfoque educativo respetuoso y relevante para la prevención, un enfoque para el que el desarrollo con IA se ha demostrado productivo en otros ámbitos (Barrera Yañez et al., 2023; Palacios & Huertas, 2024).

5.2. Efectividad del cuestionario

A pesar de haber escogido productos culturales que conocen y/o disfrutan especialmente, cabe destacar que, en general, quienes participaron en el piloto respondieron el cuestionario desde una distancia crítica. Entendemos dicha distancia como la capacidad de cualquier persona para analizar un producto cultural (Hall, 1973; Morley, 1996) y reconocer aquellas cuestiones problemáticas que contiene más allá de las propias preferencias e identificaciones (Kellner & Kim, 2010; Tortajada & Willem, 2019). El cuestionario inicial fue

formulado para detectar representaciones y justificaciones de las violencias machistas, y solo se utilizó el guión elaborado por Copilot (v. 2024) una vez que fue validado por el equipo investigador. El cuestionario generado por la IA invita a responder de forma concisa las preguntas propuestas y, a pesar de la brevedad de las reflexiones vertidas por quienes participaron en el piloto, dio pie a numerosas reflexiones sobre la representación de las violencias machistas.

Hay que insistir en que los productos culturales no fueron escogidos porque trataran sobre violencias machistas o incluyeran dichas violencias como parte del relato, sino por el gusto personal de cada participante. La herramienta, en este sentido, pretende ayudar a identificar si hay o no presencia de violencias machistas y qué tipo de representación se ofrece. Por ello, hay que tener en cuenta que se pueden producir una gran variedad de respuestas, pero todas evidencian el potencial reflexivo del ejercicio en sí de aplicar el cuestionario, ya que, en general, en las respuestas se detectan referencias varias a las opresiones de género. Una persona contesta:

> Aunque no es un tema tratado en profundidad [refiriéndose a una serie, *The Owl House*], sí que de forma indirecta habla del problema que supone el machismo sistemático (NP13:4)

Dado que son productos culturales escogidos porque han sido consumidos en algún momento (o se consumen habitualmente), no es de extrañar que las lecturas que se hacen de las series de televisión, las películas o los videojuegos sean preferentes, es decir, que compartan el punto de vista y los valores de quienes las han producido (Hall, 1973; Morley, 1996). Esto no implica una postura acrítica ya que, por una parte, siempre existe una lectura activa y, por otra, algunos de estos productos contienen mensajes que cuestionan las relaciones de poder (Morley, 1996; Kellner & Kim, 2010). Además, la propia formulación del cuestionario está pensada para incidir en los significados que se atribuyen a un producto cultural, pues, a fin de cuentas, este se elaboró para proponer temas de análisis habitualmente invisibilizados.

Muchas de las respuestas muestran que las personas participantes en el piloto se identifican tanto con sus personajes que incluso hacen suyas algunas de sus frases: 'dos pasos para delante y un paso para atrás sigue siendo un paso adelante' (D22:4). También se aprecia en ellas una detección de ciertas dimensiones simbólicas de la violencia machista –'se ve cómo algunos hombres están por encima de las mujeres cuando hablamos de poder o de toma de decisiones y realización de tareas'(H10:4)–, o de la no visibilización de las violencias machistas –'a pesar de ser una serie centrada en el sexo y en las relaciones hombre-mujer, no se tratan temas como la violencia machista; simplemente se hace una breve referencia a los micromachismos que padecen las protagonistas' (H29:4)– o de la estilización misma de la violencia (H14:9). En algunos casos, y como veremos posteriormente, las personas detectan incluso dinámicas de poder patriarcales (H14:7) o sociales y especistas (NB16:5) que la IA generativa no reconoce.

Aunque en alguna ocasión puntual se justifiquen las bromas machistas de una serie por juzgarlas inofensivas (D18:4), las respuestas del piloto muestran que el cuestionario permite activar la reflexividad y la autoconciencia, así como poner en valor la propia cultura popular y los referentes que esta genera: 'pienso que la serie te ayuda a reflexionar, no ha de ser un tutorial de lo que hacer, pero puedes coger consejos de lo que hacer si te pasa o le pasa a alguien que conoces. Además, te ayuda a ver las cosas de otra forma' (D22:6).

5.3. Coincidencia entre las respuestas de la IA y las de los y las participantes en el piloto

De la comparación de las respuestas de los cuestionarios elaborados por quienes participaron en el piloto con las que se obtuvieron a través de Gemini 1.5, se desprende que el análisis realizado por las personas y por la herramienta coincide en un alto porcentaje tanto en las formas narrativas como en los razonamientos expresados.

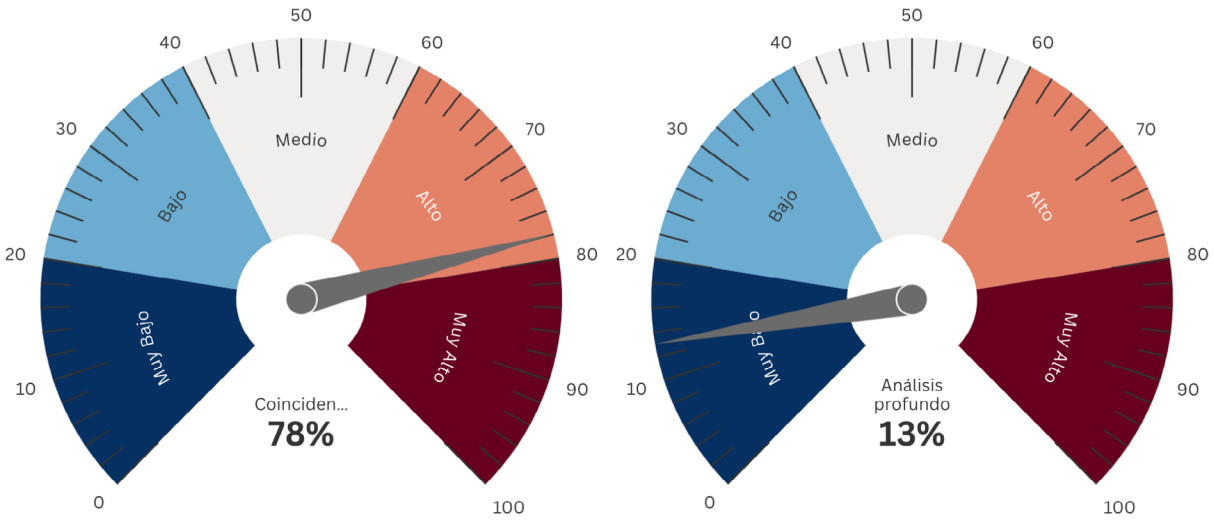

Figura 2. Coincidencias entre las respuestas de la IA y las de los participantes en el piloto. Comparación realizada con la aplicación de visualización de datos *Flourish*[3].

3 Ver https://public.flourish.studio/visualisation/21242986/

En la Figura 2 podemos apreciar el tanto por ciento de coincidencias entre la IA y las opiniones de los estudiantes (78%). Ahora bien, si comparamos los resultados de la IA con los de los estudiantes, vemos que los análisis de estos últimos son más profundos en el 13% de los casos.

Eso no significa que, en algunas ocasiones, las respuestas ofrecidas por Gemini 1.5 no respondan a un análisis más profundo y ajustado. Por ejemplo, un participante comenta que no cree que la película pretenda perpetuar o dejar de perpetuar estereotipos, y la herramienta, por su parte, responde la pregunta afirmativamente y ofrece el siguiente razonamiento para su respuesta: 'la película evita perpetuar estereotipos nocivos y presenta personajes femeninos inteligentes, capaces y complejos'.

Por una parte, esto nos hace pensar que sería posible no solo lanzar el cuestionario, sino complementarlo con otras respuestas generadas por IA para que la propia herramienta ofreciera algún tipo de retorno reflexivo, siempre acompañado por un trabajo de investigación riguroso. Para esta triangulación se requiere todavía de mucho trabajo, además de ampliar el piloto a muchos otros colectivos.

Por otra parte, no se puede obviar que la IA generativa aplica un patrón. Gemini 1.5 utiliza una serie de fórmulas para responder, tanto en lo que se refiere al contenido como al redactado empleado. En el caso del cuestionario, Gemini 1.5 aplicó tales patrones a los 28 productos culturales. Si bien la herramienta identificó en la mayor parte de las ocasiones las violencias machistas que contienen dichos productos, no siempre ofreció matices, ya que usó descripciones similares para contestar las preguntas. Además, en ocasiones usa una misma respuesta/análisis para tres series diferentes:

> Sí y no. La serie retrata con precisión las luchas y los retos a los cuales se enfrentan las supervivientes de la violencia de género, como por ejemplo el trauma y la dificultad para formar relaciones saludables. Aun así, algunos críticos argumentan que el programa también puede glamorizar o romantizar ciertos aspectos de la violencia. (Gemini 1.5)

La IA siempre ha de proporcionar una respuesta a las preguntas que se le formulan y muchas veces lo hace de forma literal, es decir, cogiendo algunas palabras clave de los enunciados del cuestionario. Trauma, salud mental, glamurización o romantización de la violencia son algunas de ellas: 'no se centra específicamente en la violencia machista, pero aborda temas de trauma y pérdida' o 'la serie retrata con precisión las luchas y los retos a los cuales se enfrentan las supervivientes de la violencia de género, como por ejemplo el trauma y la dificultad para formar relaciones saludables'.

Dado que la IA generativa domina ciertos aspectos del lenguaje, pero solo puede extraer conocimiento de datos existentes, sus respuestas se ciñen a la información disponible. El comportamiento emerge, pero la IA generativa no comprende los hechos ni puede discernir la veracidad fáctica (que no semántica) de sus afirmaciones (Nowotny, 2023).

5.4. Mayor reflexividad en las respuestas de los y las participantes

En un 13% de los casos, el participante humano hizo un análisis más profundo o realizó una detección de patrones más ajustada que la herramienta. Gemini 1.5 respondió en negativo a cuestiones en las que las personas contestaron en afirmativo. En sus respuestas, estas se refirieron, por ejemplo, a la lucha de una protagonista contra el machismo para poder desarrollar su carrera como escritora (D4:7), o a uno de los personajes que, abiertamente, denunciaba ante sus compañeros de trabajo el abuso que sufrió por parte de un cargo directivo (D22:8).

En otros casos, las personas detectan tanto las violencias simbólicas como físicas que sufren las mujeres y, en cambio, la herramienta, por falta de información, no identifica ni explicita la mediatización de estos abusos:

> Los personajes de la serie tienen sus traumas y en el caso de Haru se habla mucho de su condición de mujer y de cómo lidia con un sentimiento de inferioridad, sobre todo por su forma física. Durante la serie abusan de ella en una ocasión. Al ser una obra japonesa, se muestra de una manera no explícitamente gráfica, pero visual, que resulta poco habitual para un público occidental. Se hace duro de ver (NB16:4).
> Gemini 1.5: No.

Otros aspectos que, en ocasiones, la herramienta tampoco ha podido localizar son los que tienen que ver con la complejidad de algunos personajes femeninos en series con numerosas temporadas.

> Sí. La serie retrata un personaje femenino que intenta lidiar como puede con la manipulación de su pareja mientras busca el bienestar del resto de su familia. Aun así, el personaje ha sido criticado (creo que injustamente) por ciertas conductas (D21:8).
> Gemini 1.5: No.

Los y las participantes en el piloto han puesto en relación situaciones vividas por diferentes personajes (en este caso en una serie de televisión) y han establecido conexiones entre las experiencias de estos personajes en una reflexión global sobre los tipos de violencia que sufren las mujeres por su género o su sexualidad. Sin embargo, Gemini 1.5 puede limitarse a aplicar una fórmula:

> Sí, en varias ocasiones muestra cómo una de las protagonistas sufre violencia de género por parte de su pareja y como esta finalmente lucha para poder salir e irse con su hija. También explica cómo algunas de sus amigas son violadas por ser mujeres homosexuales (D25:4).
> Gemini 1.5: Sí y no. La serie retrata con precisión las luchas y los retos a los cuales se enfrentan las supervivientes de la violencia de género, como por ejemplo el trauma y la dificultad para formar

relaciones saludables. Aun así, algunos críticos argumentan que el programa también puede glamorizar o romantizar ciertos aspectos de la violencia.

Haber usado modelos de lenguaje y asistentes virtuales en todas las fases del proyecto nos ha permitido explorar los desarrollos de la IA en relación con las violencias machistas y, más concretamente, con su identificación y prevención. Creemos que es preferible hablar de seguimiento y monitorización que limitarnos a ser entusiastas o escépticos en general con esta tecnología y embarrarnos así en debates estériles sobre sis está a favor o en contra de la IA (Livingston, 2023; Nowotny, 2023). A fin de cuentas, como bien afirma Nowotny (2023), necesitamos desarrollos que se basen en aquello que la ciudadanía necesita, y no solo en intereses empresariales.

6. Discusión y Conclusiones

En la línea de la literatura reciente, vemos que la IA encierra numerosos sesgos de género debido a su dependencia de los datos disponibles, inherentemente sesgados a su vez (Fraile-Rojas, B. et al., 2025; Crawford & Paglen, 2019; Feast, 2019). Estos sesgos se manifiestan de diversas maneras, y es evidente que perpetúan y amplifican las desigualdades de género existentes. Por ello, creímos conveniente examinar el potencial de la IA en relación con las violencias machistas, y, concretamente, centrarnos en cómo dichas violencias están mediadas por los productos culturales, entendiendo que la propia IA se nutre de estos espacios simbólicos: prensa, críticas culturales, reseñas, etc. Asimismo, tuvimos en cuenta que la IA puede ser una herramienta para luchar contra los estereotipos y la violencia de género y que su capacidad para analizar grandes volúmenes de datos y detectar patrones ocultos ofrece novedosas oportunidades para abordar estos problemas de manera más efectiva (Palacios & Huertas, 2024).

Al poner a prueba la IA generativa elaborando una versión del test de Bechdel-Wallace centrada en la detección de las violencias machistas en la ficción, hemos podido constatar que la monitorización de la IA y su inclusión en las diferentes partes del proyecto, nos han permitido crear un instrumento para la autorreflexión pedagógicamente relevante. Si bien en algunos casos, las personas que han participado en el piloto han realizado análisis y comentarios más profundos y ajustados que los de Gemini (v1.5), la alta coincidencia entre las respuestas humanas y las de la IA abre la puerta a que este tipo de cuestionarios puedan usarse como una herramienta formativa mejorada.

Para dar más consistencia a los resultados, se han utilizado las IA generativas Copilot (v.2024) y Gemini v1.5, productos ambas de sendas compañías punteras en el desarrollo y la difusión de la IA, para explorar tanto sus posibilidades como sus limitaciones al ponerlas al servicio de un proyecto centrado en la detección y la prevención de las violencias machistas. Las dos se han demostrado útiles para tal fin guiadas, eso sí, por cierta dirección humana. En el proyecto hemos querido tener presente lo que Livingston (2023) denomina las "velocidades intermedias" de la IA. Tener en cuenta que la IA va rápida y lenta a la vez, nos ayuda a enfocarnos en aquellos espacios en los que se puede intervenir. Al detectar cómo funciona y qué aspectos del lenguaje domina la IA generativa, evitamos humanizarla y proyectar en ella nuestra propia noción de inteligencia humana; que las respuestas de las personas y del modelo multimodal sean muy parecidas solo implica que hemos identificado esta similitud. Aunque pueda parecer obvio, que el análisis de Gemini v1.5 esté formulado en términos críticos no tiene que ver con una capacidad para discernir la verdad fáctica de sus afirmaciones (Nowotny, 2023). Tampoco con la comprensión de una representación mediática y de sus implicaciones, y menos aún con la imbricación de estas experiencias mediáticas en contextos cotidianos.

Como limitaciones de esta investigación destacamos la muestra limitada a 29 sujetos humanos (28 productos audiovisuales) y el ámbito socio-geográfico localizado – un grupo natural de una asignatura en la universidad. Por lo tanto, las líneas futuras de investigación en este tipo de experimentos deberían extender la muestra, manteniendo siempre la participación de personas en el circuito para validar los resultados de la IA. Además, deberíamos generar modelos de análisis para comprender y contrarrestar las formas emergentes de abuso facilitado por la inteligencia artificial, y establecer marcos de rendición de cuentas para las decisiones impulsadas por estas herramientas.

Como reflexión final destacamos que la IA nos devuelve lo que ya se ha dicho, y solo lo que ya se ha dicho, aquí y en otras partes del mundo. Puede estar bien, pero hay que ser conscientes de sus límites, mostrar a lo que nos enfrentamos desde la ambivalencia propia de esa tecnología y preocuparnos por monitorizar su funcionamiento y su evolución. Necesitamos desarrollos que no sean tan binarios ni tan occidentales (Livingston, 2023) y que se basen en las necesidades de la ciudadanía y no solo en el interés comercial (Nowotny, 2023). Ser conscientes de los sesgos y las limitaciones que presentan estos modelos nos parece crucial para dicho desarrollo de la tecnología, pues el factor humano debería estar presente en el creciente mercado de las IA, donde cada día nos encontramos con nuevas sorpresas. Creemos que, aparte de aprender de experimentos con IA en otros campos como por ejemplo la intervención social (Guthridge et al., 2022) es importante replicar experiencias como la que se presenta en este artículo para visibilizar qué está pasando con los contenidos creados por estas herramientas con relación a la representación de las mujeres y la violencia de género en los productos de ficción.

Referencias bibliográficas

Almarcha Barbado, M. A. (2016). El oficio profesional de la sociología y otras profesiones en clave de mujer. *Investigaciones Feministas, 7*(2), 139-157. https://doi.org/10.5209/INFE.53777

Ashcraft, C., McLain, B. y Eger, E. (2016). *Women in tech: The facts*. National Center for Women & Information Technology.

Barrera Yañez, A. G., Alonso-Fernández, C. y Fernández-Manjón, B. (2023). Acceptance evaluation of a serious game to address gender stereotypes in Mexico. En *Learning technologies and systems* (pp. 229-240).

Butler, J. (1990). *El género en disputa*. Barcelona: Paidós.

Connell, M. (2002). *Gender*. Cambridge, UK: Polity; Malden, MA: Blackwell Publishers.

Crawford, K. (2021). *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*. New Haven: Yale University Press.

Crawford, K. y Paglen, T. (2019). Excavating AI: The politics of images in machine learning training sets. *AI Now Institute*.

Donnerstein, E. (1998). ¿Qué tipos de violencia hay en los medios de comunicación? El contenido de la televisión en los Estados Unidos. En J. Sanmartín Esplugues, J. S. Grisolía Thompson y S. Grisolía (Eds.), *Violencia, televisión y cine* (pp. 43-66). Barcelona: Ariel.

Feast, J. (2019). 4 ways to address gender bias in AI. *Harvard Business Review*. Recuperado de https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai

Fraile-Rojas, B., De-Pablos-Heredero, C. y Mendez-Suarez, M. (2025). Female perspectives on algorithmic bias: Implications for AI researchers and practitioners. *Management Decision,* (ahead-of-print). https://doi.org/10.1108/MD-04-2024-0884

Guthridge, M., Kirkman, M., Penovic, T. et al. (2022). Promoting gender equality: A systematic review of interventions. *Social Justice Research, 35*, 318-343. https://doi.org/10.1007/s11211-022-00398-z

Hall, S. (1973). Encoding/decoding. En *Culture, media, language: Working papers in cultural studies 1972-79* (pp. 128-138). Londres: Hutchinson.

Human Rights Watch. (2020, septiembre 29). *Automated hardship: How the tech-driven overhaul of the UK's social security system worsens poverty.* Human Rights Watch. https://www.hrw.org/report/2020/09/29/automated-hardship/how-tech-driven-overhaul-uks-social-security-system-worsens

Kaplan, A. (1998). *Las mujeres y el cine: a ambos lados de la cámara*. Madrid: Cátedra.

Kellner, D. y Kim, G. (2010). YouTube, critical pedagogy, and media activism. *Review of Education, Pedagogy and Cultural Studies, 32*(1), 3-36. https://doi.org/10.1080/10714410903482658

Livingston, S. (2023). Inteligencia artificial: en todas partes y en ninguna, rápida y lenta. En CCCB (Ed.), *IA: Inteligencia Artificial* (pp. 30-39). Barcelona: CCCB-Diputació de Barcelona.

Morley, D. (1996). *Televisión, audiencias y estudios culturales*. Buenos Aires: Amorrortu.

Mulvey, L. (1975). Visual pleasure and narrative cinema. *Screen, 16*(3), 6-18. https://doi.org/10.1093/screen/16.3.6

Nowotny, H. (2023). Parece que son como nosotros, pero no lo son. En CCCB (Ed.), *IA: Inteligencia Artificial* (pp. 40-45). Barcelona: CCCB-Diputació de Barcelona.

O'Connor, S. y Liu, H. (2024). Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities. *AI & Society, 39*, 2045-2057. https://doi.org/10.1007/s00146-023-01675-4

Palacios-Hidalgo, F. J. y Huertas-Abril, C. A. (2024). AIALL and queer language education to prevent gender-based violence: Using artificial intelligence for lesson planning. En M. Buenestado-Fernández, A. Jiménez-Millán y F. Palacios-Hidalgo (Eds.), *Comprehensive sexuality education for gender-based violence prevention* (pp. 189-210). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-2053-2.ch011

Plo-Moreno, J. (2023). Arquitectura de un sistema de ayuda a la prevención de casos de violencia de género en España. http://hdl.handle.net/10609/147237

Prates, M. O., Avelar, P. H. y Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computation & Applications, 32*(10), 6363-6381.

Schwab, K. (2016). *The fourth industrial revolution*. World Economic Forum.

Smith, G. y Rustagi, I. (2021). When good algorithms go sexist: Why and how to advance AI gender equity. *Stanford Social Innovation Review*. Recuperado de https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity

Tortajada, I. y Willem, C. (2019). Creación de significado online: recoger las voces de los y las fans de series televisivas. *Empiria, 42*, 99-112.