EDICIONES COMPLUTENSE

Investigaciones Feministas
ISSN-e: 2171-6080

MONOGRÁFICO

Agentes conversacionales inteligentes (*chatbots*) y estereotipos de género en la atención de las violencias machistas: taxonomía de posibles amenazas desde un enfoque de *threat modelling*

Borja Sanz Urquijo
Universidad de Deusto (España)
María López Belloso
Universidad de Deusto (España)
Lorea Romero Gutiérrez
Universidad de Deusto (España)
María Silvestre Cabrera
Universidad de Deusto (España)

https://dx.doi.org/10.5209/infe.100601 Recibido: Enero 2025 • Evaluado: Mayo 2025 • Aceptado: Junio 2025.

ES **Resumen: Introducción.** El uso de inteligencia artificial (IA) en la lucha contra la violencia de género (VG) ha cobrado creciente relevancia, en particular mediante agentes conversacionales o chatbots diseñados para brindar apoyo a las víctimas. Sin embargo, la implementación de estas herramientas no está exenta de riesgos. **Objetivos.** Este artículo analiza las amenazas en el uso de agentes conversacionales inteligentes o *chatbots* en la atención de la VG y propone una taxonomía de estas amenazas basada en el enfoque de *threat modeling*. **Metodología.** A partir del análisis de la literatura especializada y de experiencias existentes, se identifican vulnerabilidades tecnológicas, sesgos algorítmicos y limitaciones éticas que pueden comprometer la seguridad y efectividad de estas herramientas. Además, se discute cómo la IA Feminista puede contribuir a la creación de sistemas más inclusivos y sensibles a las necesidades de las víctimas. **Resultados y conclusiones.** La propuesta de taxonomía presentada ofrece un marco de referencia para el diseño de *chatbots* con un enfoque ético, priorizando la privacidad, la autonomía y la seguridad de las usuarias.
Palabras clave: Inteligencia artificial feminista, violencia de género, sesgos algorítmicos, chatbots, threat modeling.

ENG Intelligent conversational agents (chatbots) and gender stereotypes in addressing Gender-Based Violence: A taxonomy of potential threats from a threat modelling approach

Abstract: Introduction. The use of artificial intelligence (AI) in the fight against gender-based violence (GBV) has gained increasing relevance, particularly through conversational agents or chatbots designed to support victims. However, the implementation of these tools is not without risks. **Objectives.** This article analyzes the threats associated with the use of intelligent conversational agents or chatbots in addressing GBV and proposes a taxonomy of these threats based on the *threat modelling* approach. **Methodology.** Based on an analysis of specialized literature and existing experiences, technological vulnerabilities, algorithmic biases, and ethical limitations are identified that may compromise the security and effectiveness of these tools. Furthermore, the discussion explores how Feminist AI can contribute to the creation of more inclusive systems that are sensitive to the needs of victims. **Results and conclusions.** The proposed taxonomy provides a reference framework for designing chatbots with an ethical approach, prioritizing user privacy, autonomy, and security.
Keywords: Feminist artificial intelligence, gender-based violence, algorithmic biases, chatbots, threat modeling.

Cómo citar: Borja Sanz Urquijo, B.; María López Belloso, M.; Lorea Romero Gutiérrez, L.; María Silvestre Cabrera, M. (2025). Agentes conversacionales inteligentes (*chatbots*) y estereotipos de género en la atención de las violencias machistas: taxonomía de posibles amenazas desde un enfoque de *threat modelling*. *Investigaciones Feministas, 16*(1), 5-25. https://dx.doi.org/10.5209/infe.100601

1. Introducción

Uno de los problemas sociales y de salud pública que ha adquirido más relevancia en los últimos años ha sido el de la violencia de género (VG). Esta violencia no sólo abarca violencia física, como a menudo ha sido identificada, sino que incluye también malos tratos psicológicos y sexuales, entre otros (Krug et al., 2002). Se trata, sin duda de un fenómeno complejo y con múltiples causas que afecta a las mujeres en todas las etapas de la vida, con independencia de aspectos sociológicos como la situación económica, la religión, la profesión o su origen étnico (Jeanjot et al., 2008). Por ello, es imprescindible la creación de sistemas y servicios de apoyo a estas víctimas.

Los servicios de atención a víctimas de VG incluyen una amplia variedad de mecanismos diseñados para garantizar su protección y recuperación. Estos abarcan desde servicios telefónicos de emergencia, como el 016 en España, que brinda atención profesional las 24 horas, hasta casas de acogida que ofrecen refugio seguro para las víctimas y sus hijas e hijos. También se incluyen servicios jurídicos gratuitos, apoyo psicológico especializado, programas de inserción laboral para promover la independencia económica, y campañas de sensibilización dirigidas a la sociedad. Sin embargo, estos servicios a menudo enfrentan retos como la falta de recursos suficientes, desigualdades en su distribución territorial y barreras socioculturales que dificultan que las víctimas los utilicen. Esta situación es aún más grave en mujeres en situación de vulnerabilidad, en mujeres migradas o en mujeres en exclusión social (Toledano-Buendía, 2021).

Según el último Informe Anual del Observatorio Estatal de Violencia sobre la Mujer disponible (Observatorio Estatal de Violencia sobre la Mujer, 2024), a 31 de diciembre de 2022, el número de mujeres que habían hecho uso de los servicios de atención y protección para las víctimas de VG ascendía a 17.062, un 2,1% más que la cifra registrada en el mismo periodo del año anterior. Sin embargo, habida cuenta que, según los datos del Portal Estadístico de la Delegación del Gobierno contra la Violencia de Género, dos de cada tres víctimas no habían denunciado la situación de violencia, estamos ante un problema de una dimensión aún mayor. La magnitud de las cifras desborda la posibilidad de que los servicios disponibles ofrezcan una atención temprana y limita el número de denuncias de las víctimas que, en mayor medida, recurren antes a las redes de apoyo cercanas («Encuesta Europea de violencia de género», 2023) y, recientemente, algunas deciden denunciar en redes sociales a modo de "refugio digital" (Rekakoetxea, 2024).

Es en este contexto en el que la tecnología puede ser una herramienta eficaz. En particular, en los últimos años, el uso de tecnologías digitales y de inteligencia artificial (IA) para abordar la VG ha adquirido una atención creciente, posicionándose como una herramienta clave en la detección, prevención y apoyo a las víctimas. La digitalización de las sociedades ha ampliado las posibilidades de innovación tecnológica para enfrentar problemas sociales complejos, como la violencia de género. Sin embargo, estas tecnologías también han sido objeto de críticas desde el feminismo, por las implicancias éticas, de sesgos y de exclusión que pueden surgir en su diseño e implementación. Por ejemplo, en España se han extendido las pulseras telemáticas para monitorizar a agresores con órdenes de alejamiento, un sistema de gran relevancia pública cuyo funcionamiento (y recientes fallos) han ocupado la agenda mediática[1]. Sin embargo, estas tecnologías también han sido objeto de críticas desde el feminismo, por las implicancias éticas, de sesgos y de exclusión que pueden surgir en su diseño e implementación.

Entre todas estas tecnologías, destacan los agentes conversacionales o *chatbots*, que han emergido como una de las tecnologías más prometedoras para ofrecer apoyo a las víctimas. Estas tecnologías son sistemas de IA que están diseñadas para interactuar con las usuarias de manera automatizada, brindando información, orientación, asistencia psicológica inicial y acceso a recursos disponibles. Su capacidad para operar de forma continua, 24 horas al día, y desde cualquier lugar con acceso a internet, los posiciona como

[1] https://elpais.com/sociedad/2025-09-25/la-delegada-contra-la-violencia-de-genero-comparece-en-el-congreso-por-la-crisis-de-las-pulseras-antimaltrato.html

herramientas accesibles y discretas, especialmente valiosas para aquellas víctimas que no pueden recurrir directamente a servicios tradicionales debido a barreras de tiempo, geografía o miedo al agresor. Ejemplos como *Hello Cass*[2], un chatbot australiano diseñado para proporcionar información y apoyo en casos de VG, han demostrado el potencial de estas herramientas. No obstante, su implementación no está exenta de desafíos, como la necesidad de garantizar la privacidad de las usuarias, evitar respuestas que revictimicen y abordar los sesgos presentes en los algoritmos (Rodríguez et al., 2021).

La privacidad y seguridad de los datos de las usuarias es una preocupación central, ya que cualquier vulnerabilidad podría ponerlas en mayor riesgo. Además, los sesgos en los algoritmos pueden generar respuestas insensibles o poco inclusivas, revictimizando a las usuarias. Por otro lado, la falta de adaptabilidad cultural y las barreras de acceso, como la conectividad limitada en áreas rurales, reducen su alcance efectivo. Estos desafíos evidencian la necesidad de desarrollar *chatbots* con un enfoque ético e inclusivo, considerando tanto el diseño técnico como las necesidades psicoemocionales de las víctimas.

El presente artículo tiene como objetivo analizar el uso de tecnologías basadas en IA, específicamente estos *chatbots*, en el contexto de la VG. A través de la aplicación del enfoque de *Threat Modeling*, se busca identificar y clasificar las amenazas, vulnerabilidades y problemas éticos que surgen en estas herramientas. Además, se propone una taxonomía que permita categorizar dichas amenazas, con el fin de aportar un marco útil para quienes diseñan, legislan y desarrollan este tipo de tecnologías y muestran interés en aportar soluciones tecnológicas inclusivas y seguras para las víctimas.

El artículo está estructurado de la siguiente manera. En primer lugar, se presenta un marco teórico que explora la relación entre VG, tecnologías digitales y el enfoque crítico de la IA Feminista. Posteriormente, se analiza cómo los sesgos y estereotipos de género se manifiestan en los sistemas de IA, con énfasis en las implicancias para las herramientas destinadas a la atención de víctimas. A continuación, se revisan ejemplos de chatbots y otras tecnologías relevantes, detallando sus oportunidades y limitaciones. En la sección metodológica, se introduce el *threat modeling* y se aplica al caso de los chatbots en de asistencia a las víctimas de VG. Finalmente, se presenta una taxonomía de amenazas estructurada en cinco dimensiones principales, seguida de una discusión sobre los riesgos identificados y recomendaciones prácticas para abordar estos desafíos. El artículo concluye con un resumen de los hallazgos y propuestas para futuras investigaciones.

2. Violencia de género, IA y el enfoque de la IA Feminista

Tecnologías emergentes, como el análisis de datos, dispositivos inteligentes y aplicaciones móviles, pueden transformar la manera en que se identifica y responde a estas situaciones. Sin embargo, su implementación plantea desafíos éticos y técnicos que deben abordarse para garantizar su eficacia y equidad. A lo largo de esta sección, analizaremos distintas aplicaciones tecnológicas en este ámbito, considerando tanto sus potenciales beneficios como los riesgos asociados a su implementación, con el fin de evaluar hasta qué punto pueden contribuir a un enfoque integral de lucha contra la VG. También se introducirá la IA Feminista y el enfoque que aporta a este tipo de sistemas.

2.1. Tecnologías emergentes en la lucha contra la violencia de género

Este apartado sintetiza algunas de las aplicaciones potenciales de la tecnología recogidas por la literatura especializada sobre las distintas tecnologías que se pueden utilizar en la lucha contra la VG, y cómo estas herramientas pueden cambiar la posición de las víctimas, ayudándoles en distintas partes del proceso.

El estudio de Kouzani (2023) realiza un análisis exhaustivo de los distintos aspectos relacionados con la aplicación de la tecnología en este campo, agrupando las innovaciones tecnológicas en ocho categorías principales. En primer lugar, se destaca el análisis de datos en plataformas digitales, donde el uso de tecnologías avanzadas, como el aprendizaje profundo o *deep learning* (LeCun et al., 2015), permite procesar grandes volúmenes de información para detectar solicitudes de ayuda o situaciones de peligro a través de redes sociales. En segundo lugar, los sensores ambientales, como cámaras, micrófonos y sensores de movimiento instalados en los hogares son utilizados para identificar patrones o anomalías que podrían indicar la presencia de violencia doméstica. Otra categoría relevante es el uso de smartphones y aplicaciones, que no solo pueden detectar posibles incidentes de violencia mediante técnicas de análisis de datos, sino que también ofrecen herramientas prácticas como botones de emergencia, planes de seguridad personalizados y recopilación de pruebas legales.

Asimismo, se mencionan los dispositivos portátiles, como relojes inteligentes, que recopilan datos biométricos (frecuencia cardiaca, temperatura, entre otros) para identificar señales de estrés o agresión. La prevención del acoso no presencial es otro ámbito destacado, donde se emplea la IA para detectar y prevenir la distribución de contenido íntimo no consensuado o identificar imágenes manipuladas, como los *deepfakes* (es decir, imágenes artificiales que muestran a personas reales en escenas ficticias que parecen reales generadas con IA), que son utilizados frecuentemente como herramientas de abuso psicológico. También se incluyen las medidas anti-monitoreo, que gracias a algoritmos avanzados pueden detectar spyware, rastreadores GPS u otras herramientas empleadas para acosar o controlar a las víctimas, y alertarles sobre estos riesgos.

Por último, se señala la utilización de la realidad virtual (VR), que ofrece simulaciones diseñadas con IA para entrenar a profesionales en la respuesta a situaciones de violencia doméstica y facilitar la rehabilitación

2 https://hellocass.com.au.

de agresores a través de experiencias empáticas. Este conjunto de avances tecnológicos representa un enfoque integral para abordar, prevenir y responder a la VG mediante herramientas innovadoras. A conclusiones similares llega el estudio llevado a cabo por Rodríguez y colaboradores (2021). Ambos estudios destacan el papel crucial de la tecnología, particularmente la IA, en la detección, prevención y mitigación de la violencia. En ambos casos, se resalta cómo las herramientas tecnológicas, como los algoritmos de aprendizaje automático, los sensores inteligentes y las plataformas digitales, pueden identificar patrones de abuso y facilitar intervenciones oportunas. Además, ambos subrayan la importancia de abordar problemas éticos asociados con el uso de estas tecnologías, como la privacidad, los sesgos algorítmicos y el riesgo de mal uso. Asimismo, los dos estudios consideran a la educación como una herramienta complementaria en la lucha contra la violencia, destacando el empleo de tecnologías como la realidad virtual y los juegos serios para sensibilizar a las comunidades y capacitar a profesionales en la identificación y respuesta ante casos de abuso.

Sin embargo, Rodríguez y colaboradores (2021) realizan una revisión que identifica la IA como una de las herramientas más prometedoras en el abordaje de la VG. A diferencia de la categorización propuesta por Kouzani, estos autores agrupan las respuestas tecnológicas en cuatro categorías que solo coinciden parcialmente con las de aquel: detección offline, educación, seguridad y detección online. Entre estas, destacan que la mayoría de los estudios analizados en su revisión se centran en la detección online, representando el 50,7% de los artículos revisados (Rodríguez et al., 2021: 114625). Estos trabajos, en su mayoría, exploraban aplicaciones de IA dirigidas a identificar contenido violento en internet, subrayando la importancia de abordar no solo la violencia explícita, sino también aquellas conductas que pueden llevar a su desarrollo.

En este sentido, los artículos analizados por Rodríguez y colaboradoras abordan diversas formas de conductas relacionadas con la VG, como la misoginia, el machismo, el *grooming* infantil, la violencia entre pares y las denuncias de abuso. Según los autores, el tratamiento de estas conductas previas es fundamental para diseñar estrategias efectivas que permitan prevenir y combatir la VG en sus múltiples manifestaciones. Este enfoque amplía el alcance de la IA más allá de la identificación de incidentes, integrando su potencial para intervenir en los factores subyacentes que perpetúan la violencia. Otra de las aportaciones que Rodríguez y colaboradoras destacan es la posibilidad de detectar situaciones de VG a través del análisis de datos offline a través de técnicas de aprendizaje profundo o *deep learning*, aplicados a la identificación automática de casos de violencia a través del análisis de imágenes de lesiones (Rodríguez et al. 2021: 114628). Estos autores cuantifican en un 64% el número de estudios analizados que emplean técnicas de IA para analizar y abordar la VG.

Uno de los trabajos que más en profundidad ha analizado los retos y oportunidades de la IA aplicada a la VG es la revisión realizada por Peter Novitzky , Janine Janssen y Ben Kokkeler (2023). En su revisión sistemática estos autores parten de analizar cómo o desde qué aproximación abordan la VG los artículos que analizaron (un total de 40). Llama la atención que la mayoría de esos artículos consideraban la VG como un problema de salud pública, o sociocultural, seguido de aproximaciones más centradas en la seguridad, en la criminalidad o la justicia y siendo únicamente un 7% de todos los estudios analizados los que abordaron la cuestión como un problema de derechos humanos. Esta categorización es muy elocuente, ya que no hay apenas literatura que aborde esta cuestión desde epistemologías feministas o estudios de género.

Sin embargo, tanto esta verificación (Novitzky et al. 2023) y la analizada anteriormente (Rodríguez et al., 2021) señalan la ausencia de análisis de la VG desde la perspectiva masculina o no binaria. Es curioso que se muestre esto como una limitación y no la falta de aproximaciones feministas cuando todos estos artículos comienzan sintetizando datos que evidencian que son las mujeres y niñas las principales víctimas de este fenómeno, como reflejan las estadísticas sobre víctimas de violencia doméstica (INE, 2023). Igualmente, sorprende observar que entre las causas y origen de la VG Novitzky y colaboradoras mencionan las normas sociales, las relaciones de poder, el consumo de sustancias (drogas y alcohol) entre otras, pero no se menciona el machismo (2023: 5).

Aunque estas tecnologías ofrecen oportunidades para combatir la VG, también se ha documentado cómo los sesgos inherentes en los sistemas de IA pueden perpetuar las desigualdades estructurales. Ledesma (2022) argumenta que la IA contribuye a la "violencia simbólica" mediante la reproducción de normas culturales que refuerzan relaciones desiguales de poder, consolidando estructuras que afectan de manera desproporcionada a las mujeres y niñas. Esto subraya la necesidad de diseñar tecnologías con una perspectiva interseccional que aborde estas dinámicas.

Esta crítica pone de manifiesto la necesidad de una IA feminista, que no solo integre una perspectiva interseccional, sino que también sea capaz de cuestionar y transformar las dinámicas de poder que subyacen a la VG.

La IA feminista propone un enfoque transformador que va más allá de la mera detección de incidentes. Busca desarrollar tecnologías que no perpetúen sesgos algorítmicos ni refuercen desigualdades estructurales, sino que, por el contrario, sirvan como herramientas para desmantelar sistemas de opresión. Esto implica diseñar sistemas éticos y transparentes que incorporen voces diversas, incluidas perspectivas feministas, y que prioricen la seguridad y la dignidad de las mujeres y niñas, principales víctimas de este fenómeno. Solo a través de esta reconfiguración epistemológica y ética de la IA será posible abordar la VG de manera efectiva y sostenible.

2.2. IA Feminista: un enfoque crítico y transformador

La "IA Feminista" (FAI, por sus siglas en inglés) se presenta como un enfoque integral para abordar los sesgos de género y promover un diseño inclusivo en tecnologías de IA. Inspirada en conceptos como el

conocimiento situado de Haraway (1988) y las críticas de Alison Adam a las bases conservadoras de la IA, la FAI plantea interrogantes sobre cómo se diseñan estas herramientas, qué tipos de conocimientos se privilegian y cómo se incorporan perspectivas diversas en el desarrollo tecnológico. Este enfoque reconoce que el diseño universalista tiende a excluir a grupos marginados y propone una integración de principios interseccionales en cada etapa del desarrollo de la IA (Costanza-Chock, 2018a).

La IA feminista (FAI) puede entenderse desde varias dimensiones que abarcan su modelo, diseño, políticas, cultura y discurso crítico. En términos de modelo y diseño, Broussard (2018) la conceptualiza como un sistema que procesa datos de manera transparente y justa, incorporando principios éticos e inclusivos que aseguran la representación equitativa. Desde el ámbito de las políticas, la FAI se alinea con marcos internacionales como la Política de Asistencia Internacional Feminista de Canadá y los Objetivos de Desarrollo Sostenible de las Naciones Unidas[3], los cuales subrayan la importancia de promover la equidad en campos como la ciencia, la tecnología, la ingeniería y las matemáticas (STEM). En el aspecto cultural, ejemplos como el software Poieto[4] desarrollado por Meinders buscan incrementar la participación activa de mujeres y personas de géneros diversos en la creación de tecnologías, rompiendo barreras tradicionales en el sector. Finalmente, desde una perspectiva discursiva, la FAI se alimenta de críticas feministas y análisis de raza que examinan el impacto de la tecnología en la perpetuación de sistemas de opresión, proponiendo enfoques alternativos para su diseño y aplicación. Este enfoque multidimensional destaca cómo la FAI puede ser una herramienta transformadora en la búsqueda de justicia social y equidad tecnológica.

La FAI no solo critica los sesgos y desigualdades en las tecnologías actuales, sino que sugiere una visión transformadora sobre cómo se pueden rediseñar las herramientas de IA para ser más inclusivas, éticas y sensibles a las realidades de grupos marginados. Este enfoque se basa en la premisa de que la tecnología nunca es neutral y que las decisiones tomadas en su diseño reflejan valores y estructuras de poder preexistentes (Eubanks, 2018). Al reconocer esta relación, la FAI aboga por un rediseño que incorpore principios feministas y éticos en todas las fases de desarrollo.

La FAI expone cómo el diseño tradicional de la IA suele reproducir un modelo universalista y tecnocrático que tiende a excluir perspectivas diversas. Por ejemplo, Joy Buolamwini y Timnit Gebru (2018) demostraron que los sistemas de reconocimiento facial presentan tasas de error significativamente más altas al analizar rostros de mujeres negras, evidenciando cómo la ausencia de diversidad en los datos de entrenamiento perpetúa desigualdades. Este tipo de sesgos no solo afecta la precisión de las herramientas, sino que también contribuye a reforzar dinámicas opresivas. Frente a esto, la FAI propone prácticas que incluyen educación en STEM, cocreación comunitaria y el desarrollo de marcos éticos transparentes.

En la práctica, la FAI busca reconfigurar tanto los procesos como los resultados de la IA. Proyectos como el Feminist Data Set[5] o el Google Inclusive Images Challenge (Sculley et al., 2019) demuestran cómo los principios feministas pueden integrarse en iniciativas tecnológicas reales. Sin embargo, su implementación enfrenta desafíos debido a la resistencia institucional y las prioridades comerciales de las grandes empresas tecnológicas.

3. Sesgos y estereotipos de género en la IA

Como hemos visto en el apartado anterior, estas herramientas, que pueden ser una gran ayuda a las víctimas, tienen demostrado un doble filo: mientras que pueden facilitar la detección y la atención, también reproducen y amplifican estereotipos de género subyacentes en los datos, en el personal que los desarrolla y en los propios sistemas utilizados para su desarrollo y validación.

Son varias las investigaciones que en los últimos años han hecho hincapié en la relación entre la IA y los sesgos de género subyacentes, máxime con la aparición de los agentes de diálogo generativos (Basta et al., 2019; Kotek et al., 2023; Tang et al., 2024). Los avances de los agentes de diálogo como ChatGPT, el uso de estas herramientas está aumentando a niveles nunca vistos, (unos 100 millones de personas usuarias activas cada mes, habiendo alcanzado el millón de personas usuarias en sólo 5 días). Además, se están generando modelos similares que realizan tareas más específicas, como asistentes de compras o de atención a personas usuarias.

Estos agentes de diálogo se entrenan o refinan con grandes cantidades de datos, que suelen obtenerse principalmente de internet mediante técnicas de extracción de los datos de la web, conocidas como *web scraping*. Estos datos suelen incluir contenido tóxico, es decir, cualquier expresión, publicación o interacción que cause daño psicológico, emocional o social, ya sea de manera intencionada o no, y que contribuye a un entorno digital hostil, inseguro o excluyente (Chatzakou et al., 2019; Tahmasbi et al., s. f.).

Mientras la industria y el mundo académico siguen explorando las ventajas de utilizar el aprendizaje automático para crear mejores productos y abordar problemas importantes, los algoritmos y los conjuntos de datos en los que se basan, también pueden reflejar o reforzar percepciones y estereotipos injustos, como ocurre en el caso de la VG, siendo esto uno de los grandes problemas de los modelos grandes de lenguaje natural (Basta et al., 2019; Fast et al., s. f.; Founta et al., s. f.; Hutchinson et al., s. f.; Tan y Celis, s. f.). En la se puede observar de manera esquemática donde se pueden incorporar estos sesgos y estereotipos:

---

3 https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/priorities-priorites/policy-politique.aspx?lang=eng

4 https://www.poieto.com/

5 https://carolinesinders.com/feminist-data-set/

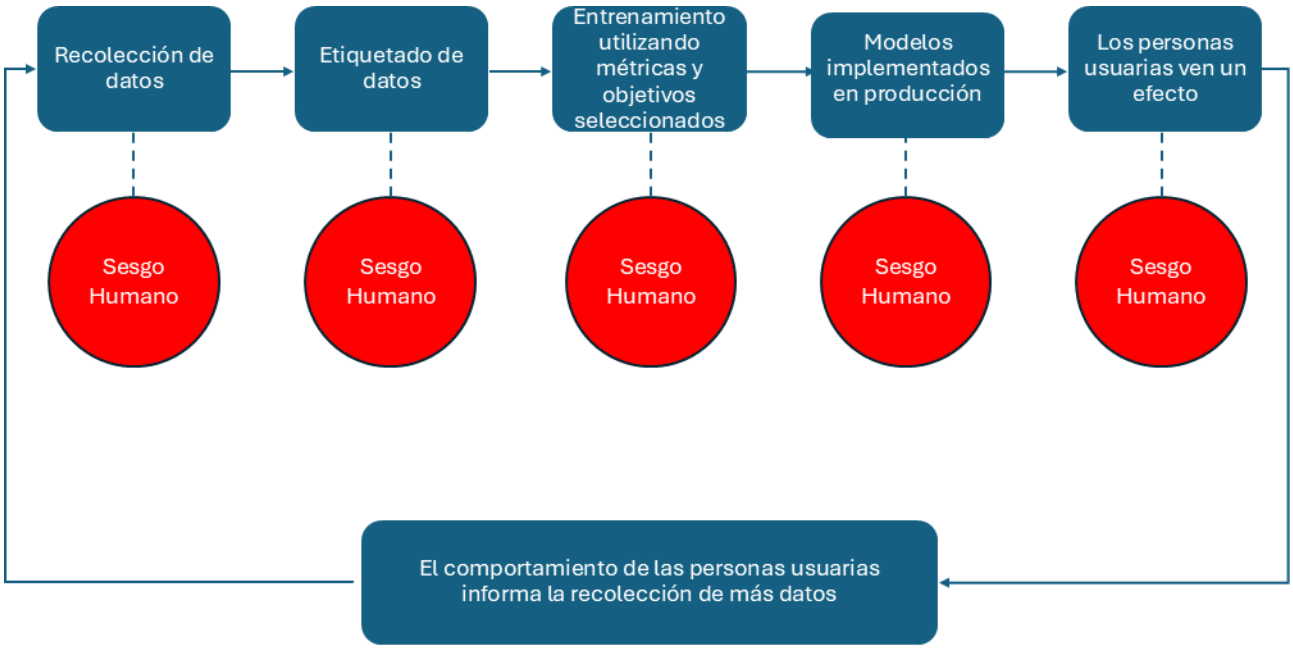

Figura 1. Esquema general de la introducción de sesgos. Fuente: Google AI Blog6. Elaboración propia.

Entre los sesgos que se pueden introducir en este tipo de sistemas, encontramos el sesgo de confirmación (tendencia a recordar información que confirma nuestras impresiones y percepciones), de automatización (tendencia a confiar más en los sistemas automatizados) o el sesgo de autoridad (confianza mayor en las opiniones de las figuras de autoridad) (Schwartz et al., 2022).

Son varios los sesgos que se pueden introducir en este tipo de sistemas. Siguiendo con lo expuesto en la Figura 1, vamos a analizar algunos ejemplos en los que se introducen estos sesgos aplicados a la prevención o abordaje de la violencia:

1. Recolección de datos: si los datos recopilados provienen principalmente de un grupo demográfico específico (por ejemplo, mujeres urbanas, nativas con acceso a internet), los datos excluirán automáticamente a mujeres rurales, migrantes o sin conexión digital. Esto crea un sesgo desde la base, ya que el sistema no refleja las experiencias de toda la población objetivo.
2. Etiquetado de los datos: si los datos se etiquetan manualmente por personas con prejuicios o conocimientos limitados, los sesgos humanos pueden introducir errores. Por ejemplo, etiquetar automáticamente frases como “estoy triste” como “no urgente” podría ignorar señales de abuso emocional, especialmente en contextos de VG, si quien establece las etiquetas no comprende la profundidad del problema.
3. Entrenamiento utilizando métricas y objetivos seleccionados: si el modelo se entrena para priorizar métricas como precisión general en lugar de identificar casos críticos, puede infrarrepresentar casos minoritarios. Por ejemplo, un *chatbot* entrenado para responder correctamente al “usuario promedio” podría no reconocer lenguaje que indique violencia en mujeres de etnias minoritarias o con vocabularios regionales específicos.
4. Modelos implementados en producción: una vez desplegado y publicado, el modelo podría priorizar respuestas rápidas en lugar de respuestas empáticas y detalladas, ya que son menos costosas económica y computacionalmente. Esto podría generar respuestas automáticas genéricas que no aborden adecuadamente las situaciones de violencia específicas, como cuando una víctima menciona aislamiento emocional o control financiero.
5. Las personas usuarias ven un efecto: si las personas usuarias reciben respuestas insensibles o inadecuadas, como “intenta hablar con alguien cercano”, pueden sentirse desatendidos o revictimizados. Esto desincentiva el uso de la tecnología y perpetúa la idea de que estos sistemas no son útiles para ciertas víctimas.
6. Ciclo de retroalimentación (comportamiento de las personas usuaria informa la recolección de datos): Si las usuarias más vulnerables dejan de usar la herramienta porque sienten que no responde a sus necesidades, los datos futuros seguirán representando únicamente a aquellas que encuentran útil el sistema, perpetuando la exclusión de las más marginadas.

Como podemos ver en la Figura 1, uno de los elementos clave para la introducción de los sesgos y los estereotipos está en el primer paso, la propia recolección de los datos. Históricamente, modelos de procesamiento de lenguaje natural como BlenderBot (Roller et al., 2021) han utilizado datos de redes sociales

6 https://research.google/blog/fairness-indicators-scalable-infrastructure-for-fair-ml-systems/

como Reddit, Twitter, 4chan, entre otras, para añadir el conocimiento a estos agentes inteligentes. Estas redes contienen un alto contenido tóxico debido al propio ambiente de la red social, que fomenta este tipo de comentarios.

A pesar de que los modelos se entrenan con grandes conjuntos de datos, no logran representar las diferentes formas en que los distintos grupos de personas ven el mundo. La participación en internet no es representativa de toda la población, por lo que los conjuntos de datos seleccionados pueden no ser los más adecuados. Por ejemplo, muchos de los conjuntos de datos se recopilan de páginas como Reddit y Wikipedia, donde la presencia de mujeres es mínima y donde la mayoría de la gente son personas jóvenes y de países desarrollados (*Demographics of Internet and Home Broadband Usage in the United States*, s. f.; Maciej Serda et al., 2020).

Además, estos conjuntos de datos suelen tener una alta presencia de ideologías supremacistas y opiniones controvertidas, al ser este tipo de temas los que más repercusión tienen en Internet y, por tanto, son aquellos que tienden a aparecer más (Bender et al., 2021). Por último, dado el enorme volumen de datos que es necesario para poder entrenar estos modelos, en la práctica, imposible comprobar y validar todos aquellos datos que se introducen en este tipo de sistemas. Este tipo de sesgos lleva a la marginalización de grupos minoritarios, que no suelen ser representados en el conjunto de datos y sus realidades no están incluidas en las respuestas de estos modelos. Como ejemplos de esta situación, podemos mencionar a TwitterBot Tay[7] y Luda[8], los cuales tuvieron que ser cerrados debido a los comentarios racistas, tóxicos y machistas que generaban.

Pese a los problemas identificados anteriormente, eso no impide que estas herramientas tengan un papel de ayuda y soporte para las víctimas de VG. El análisis realizado de la literatura indica que el uso de los agentes conversacionales y la IA como herramienta para la ayuda a las víctimas de VG puede abordarse desde dos enfoques distintos. El primero se centra en la prevención y análisis de riesgos que estas tecnologías pueden introducir sobre las mujeres que, como hemos mencionado anteriormente, pueden ser amplios (Eckstein y Danbury, 2020; PenzeyMoog y Slakoff, 2021), mientras que otro enfoque más positivista analiza las posibilidades de utilizar estas herramientas en elementos como la detección temprana o el apoyo a las mujeres que sufren VG (Kouzani, 2023; Novitzky et al., 2023; Pinto-Muñoz et al., 2023).

Por encima de estas limitaciones, la IA y los agentes conversacionales ofrecen oportunidades significativas para apoyar a las víctimas, ya sea mediante la prevención de riesgos o el fortalecimiento de redes de asistencia. En este sentido, el uso de *chatbots* y otras herramientas tecnológicas ha sido explorado como una vía para mitigar el aislamiento de las víctimas, proporcionar recursos y mejorar la seguridad, aspectos fundamentales en la lucha contra la VG.

4. Chatbots y herramientas tecnológicas en la atención a la violencia de género

El uso de la IA en la atención a las víctimas de VG se centra en la detección de situaciones de riesgo a través de dispositivos digitales. Por ejemplo, Al-Alosi (Al-Alosi, 2020) centra las principales contribuciones de la IA en 5 categorías distintas: 1. Provisión de recursos y servicios para las víctimas; 2. Mitigación de la soledad y el aislamiento; 3. Seguridad y dispositivos de protección; 4. Recolección de evidencias y 5. Empoderamiento y educación. La misma autora indica en su artículo que la VG suele conllevar aislamiento por parte de las víctimas, ya que los agresores desarrollan diversas técnicas para restringir su interacción con el entorno, lo que incluye acciones como el control de las actividades, las limitaciones de las relaciones sociales o la monitorización de las comunicaciones (Constantino et al., 2015). Estas tácticas son descritas en profundidad por la Rueda de Poder y Control de Duluth (Hasanbegovic, 2016). En esta línea, la tecnología se presenta como un elemento clave a la hora de romper ese cerco y mantener las conexiones sociales (Finn y Banach, 2000).

Entre las distintas aproximaciones sobre cómo utilizar la IA en la atención a las víctimas de VG, los agentes conversacionales prometen ser una herramienta extremadamente útil, dada su ubicuidad y su capacidad para estar permanentemente disponibles. Son varios los agentes conversacionales o *chatbots* creados para atender a las víctimas de VG a lo largo de los años. María López y Ainhoa Izaguirre (2024, p. 59 y ss.) clasifican los *chatbots* utilizados para la atención a víctimas de VG en función de sus características y objetivos. Primero, destacan los chatbots destinados a la información y empoderamiento, como *Hello Cass*[9], un *chatbot* basado en SMS que proporciona información sobre VG y orientación sobre servicios disponibles, planificación de seguridad y apoyo emocional. Asimismo, *MySis*[10] ofrece orientación práctica y apoyo emocional a través de información sobre servicios de emergencia, cuerpos policiales y tribunales, ayudando a las usuarias a tomar decisiones informadas y acceder a recursos adecuados.

En segundo lugar, mencionan los *chatbots* orientados a la intervención y seguridad, como *Sophia*[11], cuyo diseño único incluye la capacidad de borrar rastros digitales de pruebas sensibles y almacenarlas en un servidor seguro, mejorando la seguridad de las víctimas de violencia doméstica. Además, el chatbot

7 https://www.washingtonpost.com/news/the-intersect/wp/2016/03/24/the-internet-turned-tay-microsofts-fun-millennial-ai-bot-into-a-genocidal-maniac/

8 https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-facebook

9 https://hellocass.com.au/ (fecha de consulta: 07/01/2024)

10 https://changefusion.org/initiatives/11kdhvc0ebab7mgr9d85rviwj9axan (fecha de consulta 29/01/2025)

11 https://springact.org/sophia-chatbot/ (fecha de consulta: 07/01/2024)

AinoAid™[12], desarrollado en el marco del proyecto europeo IMPROVE, se enfoca en mejorar el acceso a servicios y en proporcionar asistencia personalizada a través de IA conversacional. Este *chatbot* combina el análisis de patrones de interacción con recomendaciones diseñadas para satisfacer las necesidades específicas de las víctimas.

Las autoras subrayan que, si bien estas herramientas representan un avance significativo en la lucha contra la VG, enfrentan desafíos éticos y técnicos importantes. Entre ellos, destacan los riesgos relacionados con la privacidad de los datos, el sesgo algorítmico y la necesidad de diseñar estas tecnologías desde una perspectiva interseccional y centrada en derechos humanos. Este enfoque busca garantizar que las víctimas no solo reciban apoyo eficaz, sino que también se respete su autonomía y seguridad en todo el proceso.

Para abordar esta problemática e identificar de formas más precisa las amenazas a las que se enfrentan, utilizamos el enfoque de modelado de amenazas, que permite mapear y categorizar riesgos asociados con estas tecnologías.

5. Threat Modeling como base metodológica para la creación de una taxonomía

El modelado de amenazas o *threat modelling* (Sanz et al., 2010) es un enfoque metodológico originalmente diseñado para identificar y analizar amenazas en sistemas de información, pero que también se puede adaptar para abordar problemáticas sociales complejas como las violencias machistas. Se trata de una descripción de un conjunto de aspectos relacionados con la seguridad que permite a las personas que lo desarrollan hacer un análisis global de la situación de un sistema. Este análisis prospectivo busca detectar todos los posibles puntos de fallo para después proponer contramedidas que mitiguen los problemas en caso de que estos ocurran. De esta forma, se obtienen sistemas mucho más robustos y seguros.

Esta metodología, pese a estar orientada al ámbito de la seguridad informática, permite evaluar cualquier sistema en función de sus interacciones con otros elementos, ya sean sistemas informáticos, actores humanos o entornos operativos. En este caso, su uso resulta especialmente pertinente para analizar el impacto de los *chatbots* como asistentes para víctimas de VG, dado que su implementación conlleva riesgos específicos relacionados con la privacidad, la eficacia en la respuesta y la posible revictimización de las usuarias. Aunque este estudio se ha centrado en los *chatbots*, la misma metodología podría aplicarse a otras herramientas del sistema de atención a víctimas, identificando amenazas potenciales en cada fase del proceso de asistencia. En este caso hemos centrado el modelado de amenazas en el uso de *chatbots* como asistentes para víctimas de VG, pero este análisis se podría extender a cualquier parte del sistema de atención a las víctimas de VG. Además, hemos obviado profundizar en aspectos técnicos y de desarrollo de agentes conversacionales, aunque algunos de ellos se han mencionado en la sección 1.

Esta metodología permite identificar riesgos, verificar la arquitectura de seguridad de un sistema y desarrollar contramedidas en las distintas fases del desarrollo de un sistema (Swiderski y Snyder, 2004). Por lo tanto, analizar y modelar las amenazas potenciales que enfrenta un sistema es un paso importante en el proceso de diseño de un sistema seguro. Algunas de estas amenazas pueden estar relacionadas con la propia aplicación, lo que convierte la tarea de identificar dichas amenazas en un desafío arduo, mientras que otras están relacionadas directa o indirectamente con las infraestructuras subyacentes, tecnologías o lenguajes de programación, lo que facilita la identificación y documentación de las amenazas correspondientes.

Siendo el principal objetivo del *threat modelling* proporcionar directrices útiles sobre cómo mitigar los riesgos asociados, debemos ser capaces de distinguir los elementos que corresponden a lo que se conoce como el Círculo de Riesgo (CoR, por sus siglas en inglés) (mostrado en la Figura 2).

El método se basa en analizar cinco componentes principales, representados en el CoR: activos, que son los elementos del sistema deben protegerse; amenazas, que son aquellos eventos podrían comprometer los activos; vulnerabilidades: que señalan qué debilidades en el sistema podrían ser explotadas por las amenazas; riesgos, que son las consecuencias ocurrirían si las amenazas se materializan; y las contramedidas, aquellas soluciones se pueden implementar para mitigar los riesgos.

Esta metodología permite un análisis estructurado que guía el diseño y desarrollo de sistemas tecnológicos más seguros, éticos y adaptados a las necesidades de las usuarias.

Veamos un ejemplo de cómo se desarrolla esta metodología, centrándonos únicamente en un elemento de cada categoría del Círculo de Riesgo (CoR). Consideremos un *chatbot* diseñado para ofrecer orientación y apoyo inicial a mujeres víctimas de VG.

En este caso, uno de los activos clave son los datos confidenciales proporcionados por las usuarias en sus conversaciones con el *chatbot*, como detalles sobre su situación, su ubicación o incluso sus contactos de emergencia. Estos datos son esenciales para que el *chatbot* pueda ofrecer respuestas útiles y recursos adecuados, pero al mismo tiempo representan un alto nivel de sensibilidad, ya que su exposición podría poner en peligro la seguridad de las víctimas. Una posible amenaza sería un ciberataque dirigido a la plataforma del *chatbot*, que permita a un tercero malintencionado, como el agresor, acceder a estas conversaciones privadas. Este tipo de ataque no solo comprometería la privacidad de las víctimas, sino que podría resultar en represalias directas hacia ellas si su situación queda expuesta. Esta amenaza podría aprovechar una vulnerabilidad en el sistema, como la falta de cifrado robusto durante la transmisión de datos entre el dispositivo de la usuaria y los servidores del *chatbot*. Sin este cifrado, un atacante podría interceptar las conversaciones y acceder a información sensible.

[12] https://ainoaid.fi/ (fecha de consulta: 29/01/2025)

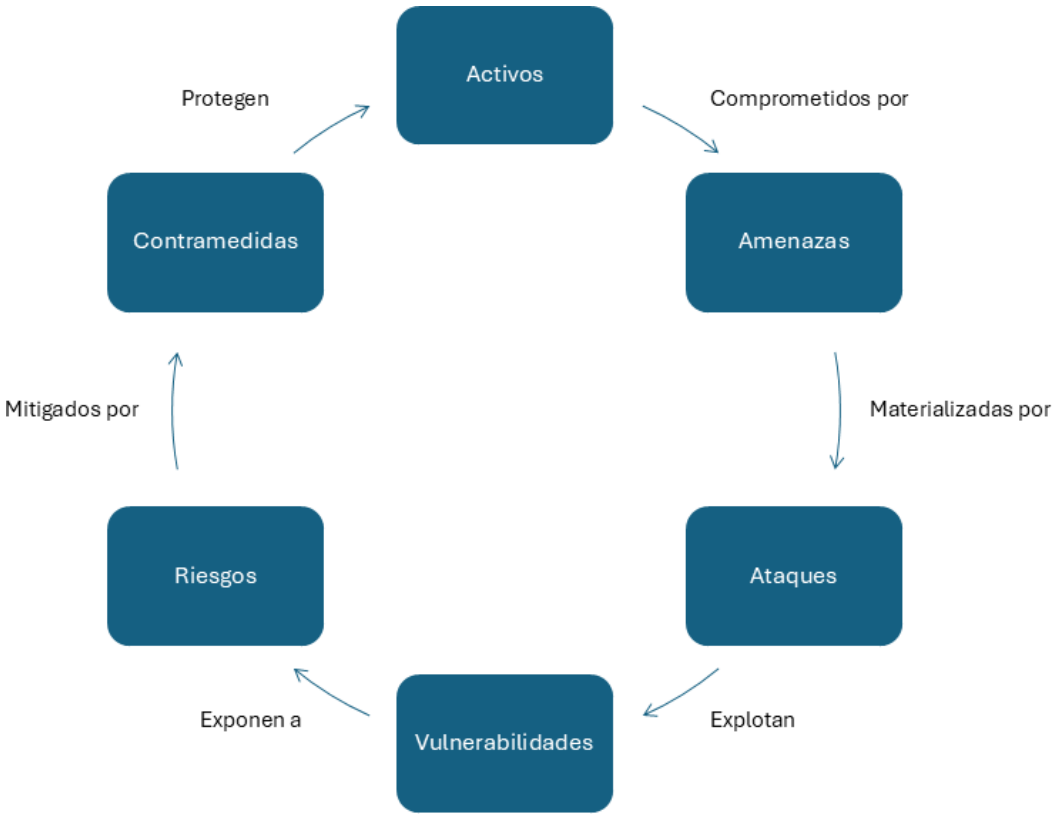

Figura 2. Círculo del riesgo del modelado de amenazas. Fuente: Elaboración propia.

De materializarse esta amenaza, el riesgo resultante sería la pérdida de privacidad de las usuarias, lo que podría llevar a situaciones graves como agresiones físicas, control más estricto por parte del agresor, o incluso la revictimización por parte de terceros. Además, este incidente dañaría la confianza de las usuarias en el *chatbot*, desincentivando su uso por parte de otras víctimas y reduciendo el impacto positivo que esta tecnología puede tener. Para mitigar este riesgo, una contramedida efectiva sería implementar un cifrado de extremo a extremo en todas las comunicaciones entre el *chatbot* y las usuarias, asegurando que los datos sean accesibles únicamente para las personas autorizadas. Además, sería esencial desarrollar políticas estrictas de manejo de datos, junto con auditorías regulares del sistema para detectar posibles vulnerabilidades. Finalmente, capacitar al equipo técnico en principios de diseño ético y sensibilidad hacia las necesidades de las víctimas garantizaría que estas herramientas sigan siendo seguras y confiables.

Este ejemplo ilustra cómo el enfoque del modelado de amenazas puede identificar riesgos específicos en herramientas tecnológicas para contextos sensibles, como la VG, y proponer soluciones para minimizarlos. Al analizar activos, amenazas, vulnerabilidades, exposiciones y contramedidas, esta metodología proporciona un marco útil para crear tecnologías más seguras y adaptadas a las necesidades de las usuarias.

El objetivo de este artículo es aplicar la metodología de *threat modeling* para mapear y categorizar los activos presentes en los *chatbots* conversacionales de asistencia a víctimas de violencia de género, así como las amenazas y vulnerabilidades asociadas con su desarrollo y uso. Sin embargo, no se busca desarrollar un modelado de amenazas completo para estos sistemas, sino ofrecer un análisis más amplio que permita comprender los riesgos desde una perspectiva integral, especialmente en el contexto de la atención y prevención de las violencias machistas. El proceso incluye varios pasos clave para abordar los riesgos asociados al uso de la inteligencia artificial. En primer lugar, se realiza una identificación de activos relevantes, los cuales, en el contexto de la IA, abarcan datos sensibles, algoritmos, interfaces y las usuarias finales. Posteriormente, se lleva a cabo un análisis de amenazas que implica evaluar los riesgos vinculados a cada uno de estos activos, considerando aspectos como sesgos algorítmicos, posibles violaciones de privacidad y el riesgo de revictimización. Finalmente, las amenazas identificadas se clasifican dentro de una taxonomía que organiza estos problemas en categorías específicas, lo que facilita su análisis y permite priorizar las acciones necesarias para mitigarlas de manera efectiva.

5.1. Aplicación del *Threat Modeling* y Círculo de Riesgo (CoR)

En esta subsección aplicamos el Círculo de Riesgo (CoR) a agentes conversacionales usados en contextos de violencia de género. El CoR distingue seis elementos que articulamos de forma homogénea en cada escenario: (a) activos a proteger, (b) amenazas previsibles, (c) vulnerabilidades del sistema, (d) riesgos para las usuarias, (e) contramedidas técnicas y organizativas y (f) una priorización que combina Probabilidad e Impacto (P×I).

Usamos definiciones operativas: Probabilidad es la frecuencia o facilidad con la que puede ocurrir el evento adverso en el contexto de despliegue; Impacto es la severidad del daño (seguridad física, privacidad,

autonomía y dignidad); la Prioridad (P×I) orienta el orden de implantación de mejoras. En todos los casos explicitamos implicaciones institucionales, gubernamentales, empresariales y comunitarias. La Tabla 1 (al final del apartado) resume los escenarios y su priorización.

5.1.1. Escenario 1. Exposición de datos sensibles por interceptación o acceso coercitivo

El primer escenario se sitúa en el plano más inmediato: proteger la conversación y sus huellas. Aun cuando el intercambio parezca inocuo, metadatos como horarios, ubicaciones aproximadas o patrones de uso pueden convertirse en señales peligrosas si un tercero (incluido un agresor) logra acceder al dispositivo, interceptar el tráfico o presionar a la usuaria para que revele información. Aquí, la mezcla de amenazas (malware, control remoto, coerción directa) con vulnerabilidades conocidas (ausencia de cifrado de extremo a extremo, historiales guardados localmente, telemetría excesiva) vuelve frágiles activos críticos: el contenido conversacional, sus metadatos y las propias reglas internas de registro.

La respuesta debe ser igualmente inmediata: comunicaciones cifradas de punta a punta, no registrar por defecto salvo que exista una necesidad proporcional y justificada, y un "modo pánico" que permita ocultar o borrar de forma discreta. Estas garantías técnicas necesitan respaldarse con pruebas independientes, con información comprensible para la usuaria sobre qué se recoge y con qué finalidad, y con procedimientos institucionales de borrado seguro. Dado que estimamos la probabilidad de incidentes en media–alta y el impacto en muy alto, la prioridad resultante es muy alta. Ello justifica, además, guías públicas de evaluación de impacto (IA y protección de datos), acuerdos de servicio con proveedores que sometan la seguridad a verificación periódica, y materiales comunitarios de "uso seguro".

5.1.2. Escenario 2. Revictimización algorítmica por malinterpretación o sesgo del LLM

Ahora bien, blindar la comunicación no resuelve por sí solo el segundo gran problema: cómo interpreta el sistema lo que la usuaria dice. Un modelo lingüístico puede leer mal las señales, minimizar el riesgo o responder sin empatía si fue entrenado con datos poco representativos o si sus indicadores de calidad premian el promedio y no los casos minoritarios. En este caso, los activos ya no son únicamente los mensajes, sino los datos de entrenamiento y evaluación, las pautas de seguridad que guían la respuesta y las rutas de derivación hacia profesionales.

La mitigación cambia de foco: se necesita curar los datos con perspectiva interseccional, integrar listas de verificación de seguridad en la generación de respuestas, y fijar umbrales conservadores que activen sin demora la derivación cuando aparezcan señales nítidas de peligro. A ello se suma supervisión humana focalizada en seguridad y la publicación de resultados desagregados por subpoblaciones, de modo que el desempeño diferencial no quede oculto por promedios. Si la probabilidad la situamos en media y el impacto en alto, la prioridad es alta. Este cálculo habilita medidas concretas: protocolos institucionales de revisión de casos críticos, estándares públicos de evaluación diferencial y compromisos empresariales de corrección con plazos, todo ello acompañado por la validación comunitaria con colectivos afectados.

5.1.3. Escenario 3. Derivación y coordinación con servicios públicos (salud, justicia, servicios sociales)

Cuando el sistema logra detectar indicios de riesgo, se abre un segundo momento igual de delicado: la derivación. En ese tránsito desde la información hacia la intervención, los activos sensibles (por ejemplo las conversaciones, identidad, localización) se conectan con plataformas públicas que operan bajo reglas y capacidades muy diferentes según el territorio. El peligro no proviene solo de fallos técnicos, sino también de errores de triaje: llegar tarde cuando había urgencia, o activar protocolos desproporcionados cuando no correspondía. Si, además, la transmisión de información se realiza sin garantías o con criterios opacos, el riesgo se multiplica.

Por ello, conviene operar con umbrales conservadores que, ante señales críticas (por ejemplo, amenazas explícitas), activen la derivación de forma automática y documentada. Esa automatización debe ir acompañada por minimización y temporalidad de los datos (no registrar por defecto, registros transitorios y cifrado punto a punto) y por acuerdos de intercambio responsable que delimiten la necesidad y la proporcionalidad de cada dato compartido. Cuando deba preservarse evidencia, esta se custodia con técnicas que eviten manipulaciones. Dado que la probabilidad varía entre media y alta según la oferta pública disponible y su madurez, y que el impacto es alto, la prioridad oscila entre alta y muy alta. En consecuencia, hacen falta protocolos interservicios con responsables y tiempos claros, auditorías periódicas con métricas desagregadas y, en paralelo, estándares gubernamentales de interoperabilidad segura que proveedores y administraciones se comprometan a cumplir y a probar conjuntamente.

5.1.4. Asistencia económica y operaciones sensibles mediadas por el chatbot

En muchos itinerarios de salida de la violencia, la dimensión económica es determinante. De ahí que algunos agentes conversacionales ofrezcan acciones como el bloqueo discreto de tarjetas, la apertura de cuentas seguras o recordatorios de ahorro de emergencia. En este entorno, los activos incluyen datos bancarios tokenizados, patrones de gasto que podrían delatar hábitos o ubicaciones, conectores con entidades de pago y preferencias de contacto. La amenaza adquiere formas muy concretas: acceso coercitivo por parte del agresor, suplantación de la tarjeta SIM, software espía que captura pantalla, o sencillamente ingeniería social dirigida a obtener claves.

La protección, en cambio, se apoya en dos principios: encubrimiento y reversibilidad. El primero se traduce en autenticación consciente de coerción (palabras o gestos que disparan silenciosamente restricciones), modos sigilosos con interfaces y notificaciones neutras, y registros efímeros que no dejen rastro innecesario. El segundo requiere topes y bloqueos temporales, así como la posibilidad de revocar de inmediato operaciones sospechosas, todo ello respaldado por cifrado extremo a extremo con la entidad financiera. Si la probabilidad es media y el impacto alto, la prioridad es alta. Este resultado obliga a implicar a bancos y proveedores en acuerdos de servicio que contemplen revocación de emergencia y pruebas periódicas, mientras que las instituciones públicas y las ONG ayudan a verificar la usabilidad real de estos mecanismos en contextos de coerción.

5.1.5. Escenario transversal. Pérdida de autonomía por sobre-automatización

Tras recorrer estos casos, emerge un denominador común: un sistema puede ser técnicamente robusto y, sin embargo, erosionar la autonomía si decisiones sensibles quedan delegadas a mecanismos opacos. La usuaria no solo necesita seguridad criptográfica o buenos modelos; necesita margen de elección y comprensión de lo que ocurre. La amenaza, en este plano, adopta la forma de automatismos que sustituyen el juicio profesional o el consentimiento informado, de empujes por defecto que orientan a una única vía, y de explicaciones insuficientes que impiden evaluar alternativas. Cuando, además, faltan controles visibles y rutas rápidas a atención humana, el resultado probable es la desconfianza: abandono de la herramienta o uso silencioso, sin activar ayudas seguras.

Por eso, cerrar el CoR muestra cómo es necesario un giro de diseño: marcar explícitamente el papel de la IA en cada interacción; ofrecer controles comprensibles para activar o desactivar automatismos y editar recomendaciones; y garantizar derivaciones a personas que no dependan de laberintos internos. Las explicaciones han de ser breves, claras y accionables, centradas en por qué se sugiere cada paso y qué opciones existen. Dado que la probabilidad de sobre-automatización la estimamos alta y el impacto medio, la prioridad sigue siendo alta. Esta conclusión se traduce en auditorías independientes de equidad y usabilidad, registros de decisiones que documenten umbrales y responsables, y métricas públicas de seguimiento (por ejemplo, el porcentaje de casos que se desvían a una persona y los tiempos de respuesta). A nivel normativo, cobra sentido reconocer el derecho a la intervención humana y exigir transparencia mínima sobre cómo operan los sistemas; en el plano empresarial, comprometer explicabilidad y accesibilidad real de los controles; y, en el comunitario, sostener una evaluación continua con usuarias y organizaciones especializadas.

La Tabla 1 sintetiza los cinco escenarios, su mapeo dentro del CoR y las valoraciones de Probabilidad, Impacto y Prioridad (P×I). En conjunto, el recorrido muestra que la seguridad no es una propiedad aislada de la tecnología, sino el resultado de decisiones de diseño, garantías organizativas y gobernanza compartida. Desde esta perspectiva, una IA feminista no solo evita daño, sino que amplía la agencia de las usuarias al convertir las protecciones en compromisos verificables por parte de instituciones, proveedores y comunidad.

Tabla 1. Tabla resumen de los distintos escenarios de ejemplo planteados.

Escenario	Activos	Amenazas	Vulnerabilidades	Riesgos	Contramedidas (clave)	Prob.	Impacto	Prioridad (P×I)
5.1.1 Exposición de datos sensibles (interceptación / coerción)1. Exposición de datos sensibles por interceptación o acceso coercitivo	Conversaciones y metadatos; claves/políticas de cifrado; reglas de logging/telemetría; canales de derivación Conversaciones, metadatos (dispositivo, timestamp)	Interceptación; acceso coercitivo; malware/control remoto del dispositivoAcceso por tercero (agresor/intruso), *malware*, interceptación	Sin E2E; historial local persistente; telemetría excesivaSin cifrado E2E; historial local; telemetría excesiva; controles de sesión débiles	Identificación de la víctima y posibles represalias/escalada de controlIdentificación de la víctima; escalada de control/coacción; represalias	E2E; **no-logging por defecto**; anonimización robusta; **modo pánico** (borrado seguro/encubrimiento); auditorías y red teaming; información clara a usuariasCifrado E2E; **no-logging** por defecto; anonimización robusta; **modo pánico**; borrado seguro; auditorías y *red-teaming*; transparencia de retención	Media–altaAlta	Muy altoAlto	**Muy altaMuy alta**
5.1.2 Revictimización algorítmica (malinterpretación / sesgo del modelo)E2. Revictimización algorítmica (LLM)2. Revictimización algorítmica por malinterpretación o sesgo del LLM	Datasets y documentación; guardrails e instrucciones; pipeline de evaluación; rutas de derivaciónLLM/PLN; *prompting*; políticas de seguridadLLM/PLN; *prompting* de sistema; políticas de seguridad	Sesgo/underfitting en subpoblaciones; débil adherencia a políticas de seguridadNo adherencia a instrucciones; sesgos en datos/objetivosNo adherencia a instrucciones; sesgos de datos/objetivos	Datos no representativos; métricas centradas en promedios; sin evaluación diferencialDatos no representativos; métricas centradas en medias; falta de evaluación diferencialDatos no representativos; métricas centradas en precisión global; falta de evaluación diferencial	Consejos inadecuados; minimización de señales; derivaciones tardíasConsejos peligrosos; tono no empático; **derivación tardía**Consejos peligrosos; tono no empático; **derivación tardía**	Curación interseccional; checklists de seguridad en generación; **umbrales conservadores** con derivación automática ante “triggers”; **supervisión humana** focalizada; reportes **desagregados**Curación interseccional; *checklists* de seguridad; **umbrales conservadores** con derivación automática; **supervisión humana**; métricas desagregadasCuración interseccional de datos; *checklists* de seguridad; umbrales conservadores con derivación inmediata; supervisión humana; reportes desagregados	MediaAltaAlta	AltoAltoAlto	**AltaMuy altaMuy alta**

Escenario	Activos	Amenazas	Vulnerabilidades	Riesgos	Contramedidas (clave)	Prob.	Impacto	Prioridad (P×I)
5.1.3 Derivación y coordinación con servicios públicosE3. Derivación a servicios públicos (salud/justicia/SS)3. Riesgos de geolocalización y trazabilidad	Datos conversacionales, identidad y localización; integraciones API con sistemas institucionalesDatos conversacionales; identidad/localización; integraciones APISistemas de ubicación; integraciones con recursos de emergencia	Errores de triaje (FN/FP); transmisión insegura; criterios opacosErrores de *triage*; filtrado inadecuado; transmisión inseguraFuga de localización; imprecisiones; exposición a terceros	Criterios de derivación opacos; señales débiles; disparidades territorialesCriterios opacos; registros persistentes; interoperabilidad débilPermisos amplios; registros persistentes; transmisión en claro	Atención inadecuada o tardía; activación de protocolos innecesariosDerivación tardía/incorrecta; exposición de datos a tercerosExposición de paraderos; fallos de derivación (ruralidad/conectividad limitada)	Umbrales conservadores; **minimización/efimeralidad** de datos; E2E punto a punto; acuerdos de intercambio responsable (necesidad/proporcionalidad); preservación segura de evidencia; **protocolos interservicios** y auditorías**No-logging** y registros efímeros; E2E; acuerdos de intercambio responsable; preservación de evidencia; protocolos interservicios**Opt-in** granular (ubicación *off* por defecto); *fuzzing* de coordenadas; minimización de retención; cláusulas de privacidad estrictas; alternativas **offline/SMS**	Media–alta (según territorio)Media–AltaMedia	AltoAltoAlto	**Alta–muy altaAlta–Muy altaAlta**
5.1.4 Asistencia económica y operaciones sensibles mediadas por chatbot 4. Transversal: pérdida de autonomía por sobre-automatización	Datos bancarios tokenizados; patrones de gasto; conectores de pago; preferencias de contacto Agencia de la usuaria; rutas de atención humana	Acceso coercitivo; **SIM-swap**; spyware/captura de pantalla; ingeniería social; exfiltración por telemetría Sustitución de juicio profesional/usuario por automatismos opacos	Autenticación no adaptada a coerción; logs persistentes; notificaciones visibles; sin “señuelo”UX opaca; ausencia de controles y vías humanas; explicabilidad limitada	Pérdida económica; trazabilidad de ayudas/refugios; escalada de control financieroDependencia tecnológica; menor capacidad de elección; abandono del sistema	Autenticación **aware-of-coercion**; **modo sigiloso** (UI/notificaciones neutras); **time-lock** y límites; **geofencing**; registros **efímeros**; E2E con entidad financiera Marcado del rol de la IA; controles/preferencias visibles; explicaciones breves; rutas rápidas a asistencia humana; co-diseño y evaluación continua	MediaAlta	AltoMedio	**AltaAlta**
5.1.5 Transversal: pérdida de autonomía por sobre-automatización	Agencia decisoria; preferencias de contacto; rutas de atención humana	Automatismos que sustituyen juicio/consentimiento; **default nudging**; explicaciones insuficientes	Falta de controles visibles; explicabilidad limitada; ausencia de vías no automatizadas	Dependencia tecnológica; reducción de opciones; desconfianza/abandono	Marcado claro del rol de la IA; **controles comprensibles** (desactivar/editar); **vías rápidas a humano**; explicaciones breves y accionables; auditorías de equidad/usabilidad; registros de decisiones	Alta	Medio	**Alta**

Elaboración propia.

5.2. Activos identificados

A partir del análisis de la literatura, hemos identificado activos clave que pueden influir en el desarrollo y efectividad de los agentes conversacionales diseñados para asistir a víctimas de VG. Estos activos se han agrupado en cuatro categorías principales: activos tecnológicos, capital humano, activos sociales y culturales, y activos legales y éticos. Cada una de estas dimensiones aporta elementos esenciales para garantizar que estas herramientas sean accesibles, seguras y culturalmente adecuadas para las usuarias.

La primera categoría, los activos tecnológicos, engloban todos aquellos elementos relacionados con la infraestructura y los sistemas que sustentan el funcionamiento de los agentes conversacionales. Un componente central son las plataformas de agentes conversacionales, que incluyen todas las aplicaciones y sistemas diseñados para interactuar de manera empática con las víctimas, brindar información personalizada y facilitar rutas de apoyo (Kouzani, 2023). Para mejorar la experiencia de las usuarias, estas plataformas se complementan con interfaces adaptativas, que permiten personalizar respuestas en función de necesidades específicas, ya sea mediante ajustes en el dispositivo, la adaptación del idioma o la inclusión de herramientas accesibles para usuarias con discapacidades (Rodríguez et al., 2021).

Otro elemento importante en esta categoría es la integración con sistemas de geolocalización, que permite a los *chatbots* detectar ubicaciones en situaciones de emergencia y conectar a las usuarias con recursos locales de asistencia (Kouzani, 2023). De manera complementaria, estos sistemas pueden enlazarse con redes de apoyo y líneas de emergencia, facilitando el acceso directo a servicios de protección o asesoramiento legal en tiempo real. No obstante, el uso de estos recursos requiere una gestión cuidadosa de la privacidad y el anonimato, razón por la cual la implementación de mecanismos de anonimización es otro activo fundamental en este contexto (Rodríguez et al., 2021). Finalmente, la incorporación de algoritmos de procesamiento de lenguaje natural permite a los *chatbots* interpretar mejor las consultas de las usuarias, detectar patrones de estrés o peligro en el lenguaje y adaptar sus respuestas de manera más comprensiva y efectiva (Rodríguez et al., 2021). Esto podría incluir, por ejemplo, adaptación de jerga usada comúnmente por las usuarias en la forma de respuesta del agente conversacional.

Más allá de la tecnología, el éxito de estas herramientas depende en gran medida del capital humano involucrado en su desarrollo y uso. En este sentido, las mujeres afectadas por la VG constituyen el grupo central de usuarias, ya que recurren a estas plataformas para denunciar, buscar apoyo y acceder a recursos de ayuda (Rodríguez et al., 2021). Sin embargo, su experiencia no solo es valiosa como destinatarias de estas tecnologías, sino también en su rol de supervivientes, ya que sus testimonios y conocimientos pueden contribuir a diseñar sistemas más inclusivos y sensibles a las distintas realidades de las víctimas (Kouzani, 2023).

Junto a ellas, las redes comunitarias y de cuidado, conformadas por familiares, amistades, colectivos feministas y ONG, desempeñan un papel clave en la asistencia a las víctimas, tanto a nivel de contención emocional como en el uso estratégico de las herramientas tecnológicas para la protección y el acompañamiento (Rodríguez et al., 2021). Además, la participación de profesionales de la psicología, el trabajo social, la abogacía y el activismo, quienes utilizan estas tecnologías para documentar casos, analizar patrones y ofrecer asistencia especializada, refuerza la efectividad de los agentes conversacionales en la atención a la VG.

Desde una perspectiva más amplia, los activos sociales y culturales desempeñan un papel esencial en la generación de conocimiento y en la consolidación de estas herramientas en los espacios comunitarios. Entre estos, destacan las narrativas y datos generados por las víctimas, los cuales permiten comprender mejor las dinámicas de la violencia y orientar el desarrollo de sistemas más contextualizados (Kouzani, 2023). De igual importancia son las comunidades virtuales de apoyo, que brindan un espacio seguro donde las víctimas y supervivientes pueden compartir experiencias, recibir apoyo emocional y aprender estrategias de seguridad (Rodríguez et al., 2021). Asimismo, las herramientas educativas, como programas de sensibilización y formación sobre VG, son fundamentales para la prevención y la transformación social (Kouzani, 2023).

Finalmente, el marco normativo y ético que regula el desarrollo y uso de estas tecnologías se recoge en la categoría de activos legales y éticos. Un punto central en este ámbito es la existencia de normas y marcos regulatorios que garantizan la protección de las víctimas y establecen estándares de uso ético de la tecnología (Rodríguez et al., 2021). Junto a estas regulaciones, las políticas de privacidad juegan un papel crucial en la protección de la información personal y sensible de las usuarias, estableciendo condiciones claras de uso y protocolos de seguridad (Kouzani, 2023). Además, se han identificado plataformas para la documentación y denuncia de incidentes de violencia, las cuales permiten a las víctimas registrar y reportar casos de manera confidencial, asegurando la trazabilidad de los hechos y facilitando el acceso a mecanismos de justicia (Rodríguez et al., 2021).

En conjunto, estos cuatro grupos de activos proporcionan un marco integral para el desarrollo de agentes conversacionales orientados a la atención de la VG. La combinación de herramientas tecnológicas avanzadas, el involucramiento de comunidades y especialistas, la incorporación de recursos culturales y la garantía de un marco normativo adecuado resulta esencial para maximizar el impacto positivo de estas soluciones y minimizar los riesgos asociados a su implementación.

5.3. Análisis de las amenazas

A partir del análisis de los activos, hemos identificado una serie de amenazas clave asociadas al uso de agentes conversacionales para el soporte a víctimas de VG. Estas amenazas se agrupan en tres dimensiones principales: tecnológicas, sociales y culturales, y legales y éticas, cada una de las cuales resalta riesgos específicos que pueden comprometer la seguridad, la efectividad y la aceptación de estas herramientas en distintos contextos de uso.

Uno de los principales desafíos tecnológicos es garantizar la seguridad de la información que manejan estos sistemas. La exposición de datos sensibles representa un riesgo crítico, ya que cualquier vulnerabilidad en el sistema podría permitir el acceso no autorizado a la información proporcionada por las usuarias. Esto podría ocurrir debido a fallos de seguridad, ataques cibernéticos o por el uso indebido de la propia tecnología, poniendo en peligro la integridad y privacidad de las víctimas. Este es un de los puntos clave que señalan Dimond y colaboradores (2011), cuando afirman:

"Es necesario desarrollar las mejores prácticas en torno al uso seguro de la tecnología y para la difusión de esta información a los defensores, el personal y las supervivientes de la violencia doméstica. También es necesario abordar la forma [...] (en la que las víctimas) ven sus habilidades técnicas en comparación con sus agresores, y cómo esto puede complicar la educación y la difusión de información sobre el uso de la tecnología."

Además, la presencia de sesgos en los algoritmos supone otra amenaza significativa, ya que puede influir en la precisión y pertinencia de las respuestas generadas por los agentes conversacionales. Como hemos visto en la sección 3, la introducción de sesgos en sistemas de IA afecta a todos los pasos en la creación de estos sistemas.

De hecho, las decisiones tomadas durante el desarrollo de estas aplicaciones pueden amplificar valores culturales y políticos arraigados, reproduciendo supuestos estereotípicos sobre las causas y soluciones de la VG. Y la solución a este problema no pasa únicamente con trabajar los conjuntos de datos, como señalan Kordzadeh y Ghasemaghaei, 2022, p.401):

> *"Los desarrolladores de algoritmos de ML y aplicaciones de IA no solo deben utilizar técnicas computacionales para mitigar los sesgos, sino también aumentar sus sistemas con transparencia, auditabilidad y funciones de control para empoderar a los usuarios para que desempeñen un papel activo en la detección y mitigación de sesgos."*

Este problema se agrava aún más con el uso de grandes modelos de lenguaje (LLMs), los cuales presentan dificultades para adherirse completamente a las instrucciones de los usuarios en diversas tareas de procesamiento del lenguaje natural. Como han demostrado (Dai y colaboradores, 2024, p. 1), estos modelos pueden malinterpretar las tareas indicadas, lo que genera desviaciones inesperadas en su desempeño:

> *"Los modelos de lenguaje de gran escala (LLMs) a menudo tienen dificultades para adherirse completamente a las instrucciones de los usuarios en diversas tareas de procesamiento del lenguaje natural [...] las desviaciones de las tareas indicadas sugieren que los LLMs pueden malinterpretar las tareas que los usuarios intentan ejecutar."*

Dado que estos modelos se entrenan con grandes volúmenes de datos, pueden reproducir estereotipos de género o generar respuestas inadecuadas que minimicen el riesgo percibido por las víctimas. Una mala calibración del algoritmo podría llevar a recomendaciones ineficaces o incluso perjudiciales, especialmente si el modelo no ha sido diseñado con una perspectiva interseccional (Broussard, 2018).

Otro problema tecnológico radica en la integración de sistemas de geolocalización, que puede presentar fallos en la identificación precisa de ubicaciones. Esto es particularmente problemático en zonas de baja conectividad o con infraestructura tecnológica limitada, como comunidades rurales, donde la inexactitud de la localización puede dificultar la conexión con servicios de emergencia (Swire et al., 2024). Además, la dependencia de la conectividad a internet es un factor de vulnerabilidad que puede dejar a las usuarias sin acceso a la herramienta en momentos críticos. En situaciones de alto riesgo, la falta de una alternativa offline o la interrupción del servicio debido a cortes en la red podrían limitar drásticamente la efectividad del *chatbot* como recurso de apoyo inmediato.

Más allá de los desafíos tecnológicos, la aceptación y apropiación de estos agentes conversacionales por parte de las usuarias también se ve influenciada por barreras sociales y culturales. Un primer obstáculo es el desconocimiento tecnológico, como señalan Dimond y colaboradores (2011): "Hemos demostrado la dificultad que deben afrontar las supervivientes de la violencia doméstica con la incorporación de las TIC: específicamente, deben lidiar con un equilibrio entre beneficios y daños". La falta de formación digital puede generar frustración y limitar el alcance de estas soluciones entre las poblaciones más vulnerables. Junto con este desafío, la desconfianza en la tecnología es otra barrera importante. El informe "Inteligencia artificial: oportunidades para la lucha contra las violencias machistas" realizado por el Equipo Deusto Valores Sociales y la Fundación InteRED y financiado en la convocatoria de BBK-kuna de 2023 (p. 50), señala:

> *"El discurso de las y los profesionales evidencia importantes inquietudes, desconfianzas e incluso rechazo ante el uso de estas aplicaciones y tecnologías, en términos de sesgos, calidad de la atención prestada y retroceso de la agencia de las mujeres en el proceso de concienciación y solicitud de ayuda."*

Algunas usuarias pueden ser reacias a interactuar con un sistema automatizado, ya sea por experiencias negativas previas con la tecnología o porque perciben la falta de empatía en la conversación con una IA. La sensación de no ser escuchadas o comprendidas puede desincentivar el uso del *chatbot* y reducir su efectividad como herramienta de apoyo (Henry et al., 2022).

Por otro lado, el estigma social asociado al uso de estas tecnologías puede actuar como una barrera adicional. En ciertos entornos, el hecho de recurrir a un *chatbot* para buscar apoyo en casos de VG puede ser malinterpretado o visto con desconfianza, exponiendo a la víctima a juicios negativos o a una mayor presión

por parte de su entorno. Esto podría generar un impacto adverso en su disposición a utilizar estos recursos digitales, afectando la eficacia de la tecnología como mecanismo de asistencia (Ngũnjiri et al., 2023).

Desde una perspectiva regulatoria y ética, existen preocupaciones importantes sobre la protección de la privacidad y la autonomía de las usuarias. Una de las amenazas más críticas es la violación de la privacidad, que puede ocurrir si las plataformas no cumplen con las condiciones de uso o si los datos recopilados son utilizados con fines comerciales o sin el consentimiento explícito de las víctimas. En estos casos, la filtración de información personal podría derivar en riesgos adicionales, como el rastreo por parte del agresor o el uso indebido de los datos en otros ámbitos (McCaughey y Cermele, 2022).

Además, la falta de regulación específica sobre el uso de IA en contextos de VG deja espacio para lagunas legales que dificultan garantizar la confidencialidad y protección de las usuarias. A pesar de avances como la IA Act (Krook, 2024), la legislación aún no cubre completamente los riesgos asociados a estas tecnologías, lo que podría generar incertidumbre en su implementación y en la rendición de cuentas en caso de fallos en el sistema. Por último, el impacto ético en la autonomía de las usuarias es un aspecto que no puede pasarse por alto. Una excesiva confianza en estos sistemas podría llevar a una reducción en la capacidad de decisión de las víctimas, delegando en la IA parte del proceso de toma de decisiones en situaciones críticas. En lugar de empoderarlas, estas herramientas podrían generar una dependencia que limite su autonomía en la gestión de su situación (Moradbakhti et al., 2022).

Si bien los agentes conversacionales representan un avance significativo en la atención y prevención de la VG, las amenazas aquí descritas evidencian la necesidad de un diseño cuidadoso y regulaciones adecuadas que minimicen sus impactos negativos. La mitigación de estas amenazas debe abordarse desde una perspectiva interdisciplinaria, integrando enfoques tecnológicos, sociales y éticos para garantizar que estas herramientas sean seguras, accesibles y realmente beneficiosas para las usuarias.

A continuación, a modo de resumen, la Tabla 1 muestra la taxonomía que relaciona los activos y las amenazas identificadas, organizadas por categoría. Este esquema permite visualizar de manera estructurada la complejidad del problema, evidenciando las interacciones entre las distintas dimensiones. La tabla ofrece un panorama integral que refuerza la necesidad de desarrollar estrategias de mitigación y diseñar herramientas inclusivas y seguras para las víctimas de VG.

Tabla 2. Tabla resumen de los activos y las amenazas identificadas, organizadas por categoría.

Categoría	Componentes o Activos	Amenazas Asociadas
Activos Tecnológicos	– Plataformas de agentes conversacionales (Kouzani, 2023). – Interfaces adaptativas (Rodríguez et al., 2021). – Sistemas de geolocalización (Kouzani, 2023). – Procesamiento de lenguaje natural (Rodríguez et al., 2021). – Integración con recursos de emergencia (Kouzani, 2023). – Mecanismos de anonimización (Rodríguez et al., 2021).	– Exposición de datos sensibles (Dimond et al., 2011) – Sesgos de algoritmos (Broussard, 2018). – Fallos en la geolocalización (Swire et al., 2024). – Interrupción de la conectividad
Capital Humano	– Mujeres afectadas por VG (Rodríguez et al., 2021). – Mujeres supervivientes (Kouzani, 2023). – Redes comunitarias (Rodríguez et al., 2021). – Agentes profesionales de intervención.	– Desconocimiento tecnológico (Henry et al., 2022) – Desconfianza en la tecnología (Henry et al., 2022) – Estigmatización (Ngũnjiri et al., 2023)
Activos Culturales y Sociales	– Narrativas y datos generados por víctimas (Kouzani, 2023). – Comunidades virtuales de apoyo (Rodríguez et al., 2021). – Herramientas educativas (Kouzani, 2023).	– Desconfianza en espacios virtuales – Estigmatización social (Ngũnjiri et al., 2023)
Activos Legales y Éticos	– Legislación aplicable (Rodríguez et al., 2021). – Políticas de privacidad víctimas (Kouzani, 2023). – Sistemas de denuncia (Rodríguez et al., 2021).	– Violación de privacidad (McCaughey y Cermele, 2022) – Falta de regulación específica (Krook, 2024), – Impacto ético en la autonomía (Moradbakhti et al., 2022)

Fuente. Elaboración propia

6. Propuestas de mitigación de las amenazas desde la IA Feminista

Ante estos desafíos, la IA feminista se plantea como un enfoque transformador que busca mitigar los sesgos algorítmicos y garantizar que estas tecnologías sean diseñadas desde una perspectiva ética e inclusiva. Desde esta mirada crítica, la IA no es una entidad neutral, ya que como señalan O'Connor y Liu (2024, p. 2046):

> *"Si bien la IA en sí misma puede ser vista como una tecnología objetiva y neutral, está imbuida de nuevos significados e implicaciones a través de su uso en contextos específicos por parte de los humanos... Como los sesgos de género son implícitos en nuestra sociedad y cultura, se convierten en parte de los 'factores contextuales' que influyen en el uso y la comprensión de las tecnologías de IA, que a su vez se ven imbuidas de los mismos sesgos".*

Este reconocimiento de la IA como una tecnología socialmente situada exige una revisión profunda del diseño algorítmico, que tradicionalmente ha operado bajo la suposición de una universalidad que ignora las diferencias estructurales entre distintos grupos de usuarios. En este sentido, la teoría del punto de vista feminista proporciona una herramienta clave para comprender que la tecnología no puede desarrollarse desde una perspectiva única y homogénea, sino que debe reconocer la pluralidad de experiencias (Dimond et al., 2011, p.420):

> *"La teoría del punto de vista feminista puede ayudar a los investigadores en HCI (Interacción Persona-Ordenador) a comprender que existe una pluralidad de experiencias [...], y que el diseño tecnológico a menudo asume un diseño "Universal" que no refleja las experiencias de todos. Específicamente, diferentes configuraciones de género, raza, clase, cultura, etc., afectan el uso de la tecnología y, por lo tanto, el diseño."*

Desde esta perspectiva, la IA feminista aboga por modelos de desarrollo que integren activamente la diversidad, garantizando que las herramientas tecnológicas no solo reconozcan las diferencias entre usuarias, sino que también se adapten a sus realidades particulares. Esto implica una transformación tanto en la recolección de datos —evitando conjuntos homogéneos que refuercen sesgos estructurales— como en las metodologías de diseño, promoviendo un enfoque participativo donde las comunidades afectadas puedan intervenir en la creación de soluciones tecnológicas que realmente respondan a sus necesidades.

Una forma de mitigar esta amenaza es mediante el empoderamiento a través del diseño participativo. Constanza-Chock (2018b) propone el Diseño de Justicia como método para empoderar a las mujeres en el proceso de diseño de la tecnología. Este método de diseño tiene como objetivo desmantelar la matriz de dominación en el diseño de tecnología y promover la liberación colectiva y la sostenibilidad ecológica. De esta manera, se busca incluir a las usuarias finales en todas las etapas del desarrollo tecnológico para que, de esta manera, se asegure que las herramientas realmente responden a sus necesidades. En esta línea, proyectos como La Fundación Vía Libre[13] hace uso de esta metodología para hacer el prototipo de una herramienta que busca reducir la discriminación en el procesamiento automático del lenguaje.

Por su parte, el uso de la automatización y de los sistemas de los agentes conversacionales puede derivar en una pérdida de la autonomía que refuerce las estructuras de poder existentes. Estas tecnologías pueden ser diseñadas para priorizar la eficiencia y la ganancia económica, sin considerar el impacto en la capacidad de decisión y la autonomía de las usuarias. Para ello es clave priorizar la autonomía en el diseño de estos *chatbots*. Los sistemas de IA deben diseñarse de manera que prioricen la capacidad de decisión y la autonomía de las usuarias, evitando que las tecnologías sustituyan su capacidad de agencia. Esto significa evitar la automatización ciega y asegurarse de que las usuarias mantengan el control sobre sus vidas y decisiones. También es necesario abordar la "optimismo cruel" (concepto desarrollado por la académica feminista Lauren Berlant que se refiere a la situación en la que las personas se apegan a objetos de deseo que, en realidad, impiden su propio bienestar o desarrollo) de las promesas tecnológicas, que pueden llevar a las personas a depender de sistemas que en realidad limitan su autonomía (Browne et al., 2023).

Es necesario replantear la noción de IA, que a menudo se basa en valores masculinos y racionalistas. En su lugar, la IA Feminista plantea explorar otras formas de inteligencia, basadas en valores cono la empatía, el cuidado y la colaboración. Esto implica repensar el propósito de la IA y cómo se puede utilizar para promover el bienestar humano y la justicia social (Browne et al., 2023, pp. 591 y ss).

Otra amenaza por reducir es la desconfianza de las mujeres en la tecnología. Desde una perspectiva feminista, se han propuesto diversas estrategias para mitigar este problema. Una de las más importantes es garantizar la transparencia y la rendición de cuentas en el desarrollo y uso de la IA. Para ello, es esencial que los procesos internos de los algoritmos sean accesibles y comprensibles, evitando la opacidad de la llamada "caja negra". Comprender cómo se toman las decisiones dentro de estos sistemas no solo fortalece la confianza de las usuarias, sino que también les permite interactuar de manera informada y crítica con la tecnología.

Además, deben existir mecanismos efectivos de rendición de cuentas que responsabilicen a quienes diseñan y despliegan tecnologías que puedan causar daño o perpetuar desigualdades. En el caso específico de los *chatbots* y agentes conversacionales, es crucial que los usuarios puedan identificarlos claramente como sistemas automatizados, asegurando que puedan distinguir entre una interacción con una IA y una

[13] https://www.vialibre.org.ar/proyecto/proyecto-diagnostico-y-mitigacion-de-sesgos-desde-america-latina/ (fecha de consulta 28/01/2025)

conversación con un profesional humano. Esta distinción no solo fomenta la toma de decisiones informada, sino que también ayuda a establecer expectativas realistas sobre el alcance y las limitaciones de estas herramientas. (Khowaja et al., 2024).

La IA feminista concibe la tecnología como socialmente situada y, por tanto, no neutral; su objetivo no es únicamente reducir sesgos, sino redistribuir agencia y responsabilidades a lo largo del ciclo de vida sociotécnico. Esta propuesta interpela explícitamente a desarrolladores y equipos de producto, a instituciones tecnológicas y académicas, a responsables de políticas públicas, a servicios de salud, justicia y acción social, a organizaciones de la sociedad civil y a la comunidad investigadora y financiadora. En respuesta a los desafíos identificados (sesgos algorítmicos, brechas de acceso, desconfianza, riesgos de privacidad y potencial de mal uso) se exige gobernanza de datos con minimización y propósito limitado, participación significativa de usuarias y profesionales en el diseño, transparencia operativa y trazabilidad con vías de recurso, evaluación interseccional con métricas por subpoblaciones, y protocolos que aseguren el escalado a atención humana en situaciones de riesgo. Asimismo, se requieren estrategias de acceso multicanal (incluidos SMS y voz), auditorías independientes y monitoreo post despliegue orientado a "métricas de daño", de modo que los agentes conversacionales funcionen como puente hacia redes de cuidado y no como sustitutos de ella.

La IA feminista ofrece un marco crítico y transformador para enfrentar los desafíos que plantean el uso de agentes conversacionales en el contexto de la VG. Como se ha señalado, la IA no puede concebirse como una entidad neutral, sino como un sistema que refleja y amplifica las estructuras de poder en las que se inserta. Por ello, resulta imprescindible repensar el propósito de los agentes conversacionales desde una perspectiva feminista, explorando enfoques alternativos basados en la empatía, el cuidado y la justicia social. Más allá de evitar la reproducción de sesgos, la IA feminista propone un cambio estructural en la forma en que se conceptualiza, diseña y utiliza la tecnología, garantizando que esta no solo sea accesible y segura, sino que también actúe como una herramienta para la transformación social.

7. Conclusiones

Los agentes conversacionales o *chatbots* han emergido como una herramienta prometedora en la atención a la violencia de género (VG), proporcionando acceso inmediato a información y asistencia. Sin embargo, su implementación conlleva desafíos significativos que deben abordarse desde una perspectiva feminista interseccional para evitar la reproducción de exclusiones y desigualdades preexistentes. Su efectividad no solo depende de su desarrollo tecnológico, sino de su capacidad para garantizar equidad, seguridad y sensibilidad a la diversidad de experiencias de las mujeres.

Uno de los principales riesgos identificados en el uso de *chatbots* es la presencia de sesgos algorítmicos, derivados de modelos de procesamiento del lenguaje natural (PLN) entrenados con datos que pueden reproducir estereotipos de género. Esto genera el peligro de ofrecer respuestas inexactas, omitir formas de violencia menos visibles o reforzar prejuicios que perpetúan la desprotección de ciertas víctimas. La IA feminista propone una reconfiguración basada en la justicia de datos y en la cocreación con las usuarias, asegurando la supervisión constante de los algoritmos para detectar y corregir estos sesgos de manera efectiva.

Otro desafío relevante es la brecha de acceso digital, que limita el impacto de estas herramientas entre mujeres en situaciones de vulnerabilidad, como aquellas con menor alfabetización digital, sin acceso a dispositivos tecnológicos o en zonas con conectividad deficiente. Ejemplos como Hello Cass, un chatbot basado en SMS que proporciona información sobre VG, demuestran la necesidad de adaptar estas herramientas a diversos niveles de acceso tecnológico. Para mitigar este problema, es fundamental implementar diseños accesibles, con interfaces inclusivas y opciones de uso en plataformas alternativas, como líneas de voz automatizadas. No obstante, los *chatbots* no deben sustituir los servicios tradicionales de apoyo, sino complementarlos, fortaleciendo redes comunitarias que garanticen el acceso a ayuda efectiva para todas las mujeres, sin distinción de su contexto socioeconómico.

La desconfianza hacia los agentes conversacionales por parte de las víctimas y profesionales de la atención a la VG representa otra amenaza clave. Muchas mujeres pueden percibir que estas herramientas carecen de la empatía necesaria para situaciones de crisis y que sus respuestas automatizadas pueden ser insensibles o inadecuadas, con el riesgo de revictimización. Además, la excesiva automatización puede afectar la autonomía de las mujeres en la toma de decisiones, generando dependencia de los sistemas tecnológicos en lugar de fortalecer su capacidad de agencia. Para contrarrestar esta problemática, la IA feminista propone el diseño de *chatbots* que prioricen la transparencia en sus respuestas, faciliten la comprensión de los procesos de toma de decisiones algorítmicos y permitan la integración con redes de apoyo humanas. De esta manera, el *chatbot* debe actuar como un puente hacia servicios de atención profesional en lugar de reemplazar la interacción humana.

Desde una perspectiva legal, el uso de *chatbots* en la atención a la VG plantea riesgos significativos en términos de privacidad y protección de datos. La recopilación y almacenamiento de información sensible sin garantías adecuadas puede exponer a las víctimas a amenazas adicionales, como la vigilancia por parte de agresores o la explotación indebida de sus datos por terceros. La ausencia de regulaciones específicas agrava esta problemática, permitiendo vacíos en la supervisión de su implementación y uso. Asimismo, la opacidad en el funcionamiento de los algoritmos puede dificultar la rendición de cuentas en casos de fallos o sesgos perjudiciales. Es imperativo desarrollar marcos normativos que establezcan directrices claras sobre el uso de la IA en contextos de VG, asegurando la protección de los derechos de las usuarias y la adopción de principios de equidad y seguridad en su diseño e implementación.

A partir de esta taxonomía, futuras líneas de investigación pueden surgir para mitigar algunas de estas amenazas. En el ámbito de las ciencias de la computación se puede explorar la optimización de los modelos de procesamiento del lenguaje natural para reducir sesgos algorítmicos y mejorar la capacidad de los *chatbots* para identificar formas sutiles de VG. Asimismo, es fundamental estudiar el impacto real de estas herramientas en la toma de decisiones de las víctimas y evaluar su integración con servicios de atención existentes. Finalmente, el desarrollo de metodologías participativas que involucren directamente a mujeres sobrevivientes de VG en el diseño y mejora de estas tecnologías podría constituir una estrategia clave para asegurar su efectividad y pertinencia.

Los agentes conversacionales aplicados a la violencia de género evidencian un potencial significativo para ampliar el acceso a información y apoyo, pero su efectividad depende de su alineación con principios de justicia, cuidado y rendición de cuentas. A la luz de la literatura revisada, su contribución será limitada si no se abordan, de forma coordinada por los actores mencionados, los sesgos en datos y modelos, la accesibilidad en contextos de baja conectividad, la transparencia necesaria para generar confianza, y las garantías robustas de privacidad y seguridad. En consecuencia, proponemos consolidar un marco de diseño y gobernanza desde la IA feminista que priorice la no sustitución del cuidado humano, la participación vinculante de usuarias y profesionales, y la evaluación continua mediante métricas de daño desagregadas; futuras líneas de investigación deberían profundizar en evaluaciones ecológicas en campo, defensas frente al mal uso (p. ej., vigilancia coercitiva y suplantación), y soluciones multicanal inclusivas, con el fin de asegurar que estas tecnologías no reproduzcan desigualdades estructurales, sino que contribuyan de manera verificable a la protección y dignidad de las víctimas.

8. Referencias bibliográficas

Al-Alosi, H. (2020). Fighting fire with fire: Exploring the potential of technology to help victims combat intimate partner violence. *Aggression and Violent Behavior*, *52*, 101376.

Basta, C., Costa-Jussà, M. R., y Casas, N. (2019). *Evaluating the Underlying Gender Bias in Contextualized Word Embeddings*. http://data.statmt.org/

Bender, E. M., Gebru, T., McMillan-Major, A., y Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery:*, 14.

Broussard, M. (2018). *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press.

Browne, J., Cave, S., Drage, E., McInerney, K., Browne, J., Cave, S., Drage, E., y McInerney, K. (Eds.). (2023). *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*. Oxford University Press.

Buolamwini, J., y Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, *77*-91. https://proceedings.mlr.press/v81/buolamwini18a.html

Chatzakou, D., Leontiadis, I., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A., y Kourtellis, N. (2019). Detecting Cyberbullying and Cyberaggression in Social Media. *ACM Transactions on the Web*, *13*(3). https://doi.org/10.1145/3343484

Constantino, R. E., Braxter, B., Ren, D., Burroughs, J. D., Doswell, W. M., Wu, L., y Greene, W. B. (2015). Comparing Online with Face-to-Face HELPP Intervention in Women Experiencing Intimate Partner Violence. *Issues in Mental Health Nursing*, *36*(6), 430-438. https://doi.org/10.3109/01612840.2014.991049

Costanza-Chock, S. (2018a). Design justice, AI, and escape from the matrix of domination. *Journal of Design and Science*, *3*(5), 1-14.

Costanza-Chock, S. (2018b). *Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice* (SSRN Scholarly Paper No. 3189696). Social Science Research Network. https://papers.ssrn.com/abstract=3189696

Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., y Xu, J. (2024). Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6437-6447. https://doi.org/10.1145/3637528.3671458

Demographics of Internet and Home Broadband Usage in the United States. (s. f.). Pew Research Center. Recuperado 6 de mayo de 2023, de https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

Dimond, J. P., Fiesler, C., y Bruckman, A. S. (2011). Domestic violence and information communication technologies. *Interacting with Computers*, *23*(5), 413-421. Interacting with Computers. https://doi.org/10.1016/j.intcom.2011.04.006

Eckstein, J. J., y Danbury, C. (2020). What is violence now?: A grounded theory approach to conceptualizing technology-mediated abuse (TMA) as spatial and participatory. *The Electronic Journal of Communication*, *29*(3-4).

Encuesta Europea de violencia de género. (2023, noviembre 8). *Delegación del Gobierno contra la Violencia de Género*. https://violenciagenero.igualdad.gob.es/violenciaencifras/encuesta_europea/

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press. https://books.google.com/books?hl=es&lr=&id=pn4pDwAAQBAJ&oi=fnd&pg=PP10&dq=Automating+Inequality:+How+High-Tech+Tools+Profile,+Police,+and+Punish+the+Poor&ots=gFRDbgssk&sig=adn0nsBn4PahUSXeERX-ZjgkwLs

Fast, E., Vachovsky, T., y Bernstein, M. S. (s. f.). *Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community*. Recuperado 6 de mayo de 2023, de www.aaai.org

Finn, J., y Banach, M. (2000). Victimization Online: The Downside of Seeking Human Services for Women on the Internet. *CyberPsychology & Behavior*, *3*(5).

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., y Stringhini, G. (s. f.). *Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior*. Recuperado 6 de mayo de 2023, de www.aaai.org

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575-599. https://doi.org/10.2307/3178066

Hasanbegovic, C. (2016). Violencia basada en el género y el rol del Poder Judicial. *Revista de la Facultad de Derecho*, *40*, 119-158.

Henry, N., Vasil, S., Flynn, A., Kellard, K., y Mortreux, C. (2022). Technology-Facilitated Domestic Violence Against Immigrant and Refugee Women: A Qualitative Study. *Journal of Interpersonal Violence*, *37*(13-14), NP12634-NP12660. https://doi.org/10.1177/08862605211001465

Hutchinson, B., Prabhakaran, V., Denton, E., y Webster, K. (s. f.). *Social Biases in NLP Models as Barriers for Persons with Disabilities*. Recuperado 6 de mayo de 2023, de https://www.worldbank.org/en/topic/disability

Jeanjot, I., Barlow, P., y Rozenberg, S. (2008). Domestic violence during pregnancy: Survey of patients and healthcare providers. *Journal of Women's Health*, *17*, 557-567. https://doi.org/10.1089/jwh.2007.0639

Khowaja, S. A., Khuwaja, P., Dev, K., Wang, W., y Nkenyereye, L. (2024). ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation: A Review. *Cognitive Computation*, *16*(5), 2528-2550. https://doi.org/10.1007/s12559-024-10285-1

Kordzadeh, N., y Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388-409. https://doi.org/10.1080/0960085X.2021.1927212

Kotek, H., Dockum, R., y Sun, D. (2023). Gender bias and stereotypes in Large Language Models. *Proceedings of The ACM Collective Intelligence Conference*, 12-24. https://doi.org/10.1145/3582269.3615599

Kouzani, A. Z. (2023). Technological innovations for tackling domestic violence. *IEEE Access*.

Krook, J. (2024). *Manipulation and the AI Act: Large Language Model Chatbots and the Danger of Mirrors* (SSRN Scholarly Paper No. 4719835). Social Science Research Network. https://doi.org/10.2139/ssrn.4719835

Krug, E., Dalhberg, L., Mercy, J., Zwi, A., y Lozano, R. (2002). *Word report on violence and health*.

LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), Article 7553. https://doi.org/10.1038/nature14539

Ledesma, J. O. (2022). Algoritmos y género: Inteligencia artificial al servicio de la violencia simbólica. *Revista Llapanchikpaq: Justicia*, *4*(5), 209-236. https://doi.org/10.51197/lj.v4i5.659

López Belloso, M., y Izaguirre Choperena, A. (2024). Nuevas formas de atención a situaciones de violencia de género: La irrupción de la inteligencia artificial en la atención a las mujeres víctimas. *La protección de las víctimas de la violencia de género: aspectos jurídicos y asistenciales, 2024, ISBN 9788413252377, págs. 47-86*, 47-86. https://dialnet.unirioja.es/servlet/articulo?codigo=9920844

Maciej Serda, Becker, F. G., y Cleary, M. (2020). Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia. *Uniwersytet Śląski*, *7*(1), 343-354. https://doi.org/10.2/JQUERY.MIN.JS

McCaughey, M., y Cermele, J. (2022). Violations of Sexual and Information Privacy: Understanding Dataraid in a (Cyber)Rape Culture. *Violence Against Women*, *28*(15-16), 3955-3976. https://doi.org/10.1177/10778012211070316

Moradbakhti, L., Schreibelmayr, S., y Mara, M. (2022). Do Men Have No Need for "Feminist" Artificial Intelligence? Agentic and Gendered Voice Assistants in the Light of Basic Psychological Needs. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.855091

Ngũnjiri, A., Memiah, P., Kimathi, R., Wagner, F. A., Ikahu, A., Omanga, E., Kweyu, E., Ngunu, C., y Otiso, L. (2023). Utilizing User Preferences in Designing the AGILE (Accelerating Access to Gender-Based Violence Information and Services Leveraging on Technology Enhanced) Chatbot. *International Journal of Environmental Research and Public Health*, *20*(21), Article 21. https://doi.org/10.3390/ijerph20217018

Novitzky, P., Janssen, J., y Kokkeler, B. (2023). A systematic review of ethical challenges and opportunities of addressing domestic violence with AI-technologies and online tools. *Heliyon*.

Observatorio Estatal de Violencia sobre la Mujer. (2024). *XVI Informe Anual del Observatorio Estatal de Violencia sobre la Mujer 2022* (p. 522). Ministerio de Igualdad. Centro de Publicaciones.

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

PenzeyMoog, E., y Slakoff, D. C. (2021). *As technology evolves, so does domestic violence: Modern-day tech abuse and possible solutions*. Emerald Publishing Limited.

Pinto-Muñoz, C. C., Zuñiga-Samboni, J. A., y Ordoñez-Erazo, H. A. (2023). Machine Learning Applied to Gender Violence: A Systematic Mapping Study. *Revista Facultad de Ingeniería*, *32*(64).

Rekakoetxea, Z. (2024). Refugio digital y fracaso judicial: La denuncia de la violencia machista. *Newsletter de la Asociación Vasca de Sociología y Ciencia Política*.

Rodríguez, D. A., Díaz-Ramírez, A., Miranda-Vega, J. E., Trujillo, L., y Mejía-Alvarez, P. (2021). A Systematic Review of Computer Science Solutions for Addressing Violence Against Women and Children. *IEEE Access*, *9*, 114622-114639. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3103459

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., y Liu, Y. (2021). Recipes for Building an Open-Domain Chatbot. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics*, 300-325. https://doi.org/10.18653/v1/2021.eacl-main.24

Sanz, B., Laorden, C., Alvarez, G., y Bringas, P. G. (2010). A Threat Model Approach to Attacks and Countermeasures in On-line Social Networks. *11th Reunion Española de Criptografía y Seguridad de la Información (RECSI)*, 343-348. http://paginaspersonales.deusto.es/bosanz/publications/pdf/2010/sanz_RECSI10_A%20Threat%20Model%20Approach%20to%20Attacks%20and%20Countermeasures%20in%20OSN.pdf

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., y Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (No. NIST SP 1270; p. NIST SP 1270). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.SP.1270

Sculley, D., Breck, E., Ivanov, I., Atwood, J., Skalic, M., Baljekar, P., Ostyakov, P., Solovyev, R., Wang, W., y Halpern, Y. (2019). *The inclusive images competition*.

Swiderski, F., y Snyder, W. (2004). *Threat Modeling*. Microsoft Press.

Swire, P., Kennedy-Mayo, D., Bagley, D., Krasser, S., Modak, A., y Bausewein, C. (2024). Risks to cybersecurity from data localization, organized by techniques, tactics, and procedures. *Journal of Cyber Policy*, *9*(1), 20-51. https://doi.org/10.1080/23738871.2024.2384724

Tahmasbi, F., Schild, L., Ling, C., Blackburn, J., Stringhini, G., Zhang, Y., y Zannettou, S. (s. f.). *"Go eat a bat, Chang!": On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19*. https://doi.org/10.1145/3442381.3450024

Tan, Y. C., y Celis, L. E. (s. f.). *Assessing Social and Intersectional Biases in Contextualized Word Representations*. Recuperado 6 de mayo de 2023, de https://github.com/openai/gpt-2-output-dataset

Tang, K., Zhou, W., Zhang, J., Liu, A., Deng, G., Li, S., Qi, P., Zhang, W., Zhang, T., y Yu, N. (2024). GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models. *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 1196-1210. https://doi.org/10.1145/3658644.3670284

Toledano-Buendía, C. (2021). Barrera lingüística y victimización secundaria: La (des)atención institucional a las víctimas extranjeras de violencia de género en España. *Verba Hispánica*, *29*, 175-191.