OBD RETIS, 2025, 2 , 31-41
https://doi.org/10.70664/retis.v2i1.004

REVISTA DE **TECNOLOGÍA** PARA LA **INCLUSIÓN SOCIAL**

ORIGINALES

La interseccionalidad como herramienta analítica para la auditoría de sesgos en inteligencia artificial y modelos de lenguaje a gran escala (LLM)

Intersectionality as an analytical tool for bias auditing in artificial intelligence and large language models (LLM)

Jose María Regalado López
Universidad Pontificia de Comillas

PALABRAS CLAVE:
Interseccionalidad
auditoría algorítmica
sesgo algorítmico
inteligencia artificial
equidad

RESUMEN:
El objetivo de este artículo es postular la necesidad urgente de adoptar un marco de análisis interseccional para identificar y comprender las formas complejas de sesgo que emergen en la inteligencia artificial (IA) y los modelos de lenguaje a gran escala (LLM). la metodología examina la arquitectura de la inequidad algorítmica a lo largo del ciclo de vida de la IA (datos, modelado e implementación), utilizando la teoría interseccional como herramienta analítica crítica y describiendo una metodología en tres fases para la auditoría: mapeo del contexto situado, desagregación de métricas de equidad y comunicación de la complejidad. se discute que los análisis unidimensionales son insuficientes, pues la discriminación es compuesta y afecta de forma más grave a grupos en la intersección de identidades. Se concluye que la interseccionalidad es una herramienta esencial para una auditoría de sesgos efectiva que aspire a la justicia social, requiriendo un enfoque interdisciplinario para garantizar sistemas de IA equitativos y confiables.

KEYWORDS:
Intersectionality
algorithmic auditing
algorithmic bias
artificial intelligence
equity

ABSTRACT:
The objective of this article is to postulate the urgent need to adopt an intersectional analysis framework to identify and understand the complex forms of bias that emerge in Artificial Intelligence (AI) and Large Language Models (LLM). The methodology examines the architecture of algorithmic inequity throughout the AI lifecycle (data, modeling, and implementation), utilizing intersectional theory as a critical analytical tool and describing a three-phase methodology for the audit: mapping of the situated context, disaggregation of equity metrics, and communication of complexity. It is discussed that unidimensional analyses are insufficient, since discrimination is composite and affects groups at the intersection of identities more severely. It is concluded that intersectionality is an essential analytical tool for effective bias auditing that aspires to social justice, requiring an interdisciplinary approach to guarantee equitable and reliable AI systems.

1. INTRODUCCIÓN

El auge de los Modelos de Lenguaje a Gran Escala (LLM), como GPT-4 de OpenAI y LLaMA 2 de Meta, marca un punto de inflexión en la integración de la Inteligencia Artificial (IA) en la sociedad. Estos sistemas están siendo rápidamente incorporados en ámbitos sociales, económicos y culturales, desde la generación de contenido hasta la toma de decisiones críticas en áreas como la salud, el derecho y la economía. Sin embargo, a pesar de su impresionante potencial, estos sistemas no son neutrales.

Los LLM son entrenados con vastos *corpus* de texto extraídos de internet, lo que inevitablemente provoca que absorban, reproduzcan e incluso amplifiquen los sesgos y estereotipos presentes en la sociedad y en los datos históricos. Esta incorporación de prejuicios inherentes en los datos de entrenamiento constituye una de las principales fuentes de sesgo algorítmico (Navigli et al., 2023; Ovalle et al., 2023).

Las consecuencias negativas de esta realidad son palpables y tienen un impacto directo en la justicia y la equidad social. Los sesgos se manifiestan en forma de daños asignativos, donde se distribuyen recursos u oportunidades de manera injusta (Ovalle et

CÓMO CITAR: Regalado, J. M. (2025). La interseccionalidad como herramienta analítica para la auditoría de sesgos en inteligencia artificial y modelos de lenguaje a gran escala (LLM). *RETIS. Revista de Tecnología para la Inclusión Social, 2(1), 31-41,* DOI:10.70664/retis.v2i1.004

* C-e: direccion@josemariaregalado.com

Manuscrito Aceptado: 05/10/2025

al., 2023). Un ejemplo notorio es el caso del software COMPAS, utilizado en el sistema judicial de EE.UU., donde se documentó que el algoritmo perpetuaba la discriminación racial al predecir incorrectamente una mayor tasa de reincidencia para acusados negros en comparación con los blancos. De igual manera, se ha evidenciado el sesgo de género en sistemas de selección de personal, como el caso de Amazon, que desechó un algoritmo de contratación al descubrir que éste favorecía consistentemente a los hombres basándose en patrones lingüísticos sesgados (Costanza-Chock et al., 2022). Estos modelos también generan daños representacionales, perpetuando estereotipos, por ejemplo, al asociar roles estereotípicos de género como "programador" con el hombre y "ama de casa" con la mujer (Kotek et al., 2023).

Ante esta complejidad, los análisis de sesgo unidimensionales, que se centran en un único eje de identidad (como el género o la raza), han demostrado ser insuficientes para capturar la naturaleza compuesta y entrelazada de la discriminación algorítmica (Ovalle et al., 2023; Simó, 2024). La desigualdad social está determinada por múltiples ejes (raza, género, clase) que actúan de manera conjunta y se influencian entre sí (Hill Collins y Bilge, 2019).

Por lo tanto, este artículo postula la necesidad urgente de adoptar un marco de análisis interseccional para identificar y comprender las formas complejas y compuestas de sesgo que emergen en la intersección de múltiples identidades. El enfoque interseccional, acuñado originalmente por Kimberlé Williams Crenshaw (1990), ofrece la herramienta analítica crítica para desvelar la Matriz de Dominación que subyace a la exclusión (Crenshaw, 1989; Hill Collins, 2017; Ovalle et al., 2023).

Describimos unas pautas metodológicas para integrar sistemáticamente la perspectiva interseccional en las prácticas de auditoría algorítmica, transformando el proceso de una simple verificación técnica a un análisis sociotécnico robusto (Raji et al., 2020). Al adoptar la auditoría de sesgo en conjunto con el pensamiento complejo, se garantiza que los sistemas de IA no solo sean técnicamente robustos, sino también deseables y éticamente alineados con los principios de justicia y equidad social (Bustelo, 2024).

Este artículo se estructura de la siguiente manera: primero, se examina cómo los sesgos sociales, como el sesgo de género y el sesgo racial, se introducen y se propagan a lo largo de todo el ciclo de vida de los LLM. A continuación, se profundizará en la teoría interseccional como una herramienta analítica crítica para superar las deficiencias de los enfoques de eje único. Posteriormente, se propondrá un modelo práctico para integrar el análisis interseccional en la metodología de las auditorías algorítmicas, un mecanismo de gobernanza clave para la rendición de cuentas (Eticas Research and Consulting, 2021). Finalmente, se discutirán los desafíos culturales, lingüísticos y metodológicos que esta integración plantea.

2. LA ARQUITECTURA DE LA INEQUIDAD ALGORÍTMICA. UNA TAXONOMÍA DE SESGOS A LO LARGO DEL CICLO DE VIDA DE LA INTELIGENCIA ARTIFICIAL

La creciente integración de la Inteligencia Artificial en sistemas sociales y económicos críticos ha puesto de manifiesto la inevitable transferencia y amplificación de los prejuicios humanos y las desigualdades sociales históricas dentro de los sistemas algorítmicos. Eliminar este fenómeno, conocido como sesgo de la IA o sesgo de *machine learning*, es un reto no solo técnico, sino profundamente social, ya que requiere un entendimiento profundo de las fuerzas sociales existentes y de las técnicas de la ciencia de datos. Los sesgos pueden encontrarse en cada etapa del ciclo de vida de la IA.

El análisis de estos sesgos en los LLM debe ser una tarea estratégica que abarque todas las fases de su ciclo de vida, desde la concepción y la recopilación de datos hasta su implementación y evaluación. Los sesgos no son un problema aislado que se corrige en una única etapa, sino un fenómeno sistémico que puede ser introducido, reforzado o amplificado en múltiples puntos del proceso de desarrollo.

A continuación, se presenta una taxonomía del sesgo algorítmico estructurada en tres fases principales a lo largo del desarrollo y la implementación del sistema.

2.1. FASE 1: SESGOS PRE-ALGORÍTMICOS EN LA PREPARACIÓN Y DATOS

El sesgo pre-algorítmico surge antes de la construcción del modelo, principalmente en la selección, recopilación y etiquetado de los datos utilizados para entrenar los modelos de *machine learning*[1]. Puesto que los sistemas de IA aprenden a tomar decisiones basándose en estos datos, es crucial que los conjuntos de datos sean evaluados para detectar la presencia de sesgos. Si la calidad del dato es deficiente, la calidad del resultado del algoritmo se verá comprometida. (Navigli et al., 2023).

2.1.1.Fuentes fundamentales del sesgo de datos

El problema central radica en que los datos de entrenamiento a menudo reflejan y perpetúan los prejuicios sociales preexistentes.

Representación desigual o sesgo de muestreo

Este es un factor crítico. Se produce cuando ciertos grupos están sobrerrepresentados o infrarrepresentados en el conjunto de datos de entrenamiento con respecto a la población real a la que se aplicará el sistema. (Chu et al., 2024).

Un ejemplo notable es el de los algoritmos de reconocimiento facial, donde la sobrerrepresentación de personas de raza blanca en los datos de entrenamiento puede llevar a errores significativos en el reconocimiento de personas de color. De manera similar, se observaron disparidades en conjuntos de datos como IJB-A y Adience[2], donde los individuos de piel clara constituían una vasta mayoría, sesgando el análisis contra grupos subrepresentados de piel oscura.

1 El Aprendizaje de Máquinas (Machine Learning) es un subcampo de la Inteligencia Artificial (IA) que utiliza algoritmos y técnicas de aprendizaje automático para que los sistemas puedan aprender a tomar decisiones o detectar patrones y reglas latentes directamente a partir de la ingestión de datos de entrenamiento.

2 Los conjuntos de datos IJB-A y Adience son bancos de pruebas (benchmarks) de imágenes faciales no restringidas utilizados en visión por computadora para tareas de clasificación y reconocimiento, y son notables por exhibir un sesgo demográfico significativo con una sobrerrepresentación de hombres de piel clara y una infrarrepresentación de mujeres de piel oscura

Sesgo de etiquetado humano
Los prejuicios inherentes en las percepciones o juicios de los etiquetadores (o anotadores) pueden manifestarse en los datos de entrenamiento, perpetuando estereotipos y decisiones discriminatorias (Crawford, como se citó en Bustelo, 2024), El uso de etiquetado incoherente o la inclusión/exclusión desproporcionada de ciertas características en los datos pueden llevar a la exclusión de solicitantes de empleo cualificados, por ejemplo (IBM Data and AI Team, 2023).
Sesgo histórico y social
La realidad capturada en los datos históricos a menudo está marcada por la injusticia y la desigualdad (por ejemplo, si los datos muestran que "todos los programadores son hombres y todas las enfermeras son mujeres," el modelo interiorizará y perpetuará estos sesgos ocupacionales y de género) (Chu et al., 2024). El vocabulario y los indicadores socioeconómicos (como los ingresos) utilizados por el algoritmo pueden discriminar involuntariamente por raza o sexo.

2.1.2.Tipos de sesgos sociales incorporados

Los modelos de lenguaje grande (LLM) y otros sistemas de IA demuestran incorporar una amplia gama de sesgos sociales a partir de sus datos de entrenamiento, afectando especialmente a las minorías y los grupos marginados (*Navigli et al., 2023),* los más comunes son:

- **Sesgo de género:** Es uno de los más estudiados. Los sistemas asocian roles estereotípicos (como "programador" para hombre y "ama de casa" para mujer) (Kotek et al., 2023). Este sesgo se ha cuantificado en incrustaciones de palabras, revelando estereotipos étnicos y de género. Estudios sobre ChatGPT han encontrado sesgos de género en documentos generados, como cartas de referencia, donde los textos para sujetos masculinos muestran una formalidad, positividad y capacidad de acción significativamente mayores que los de sujetos femeninos (Wan et al., 2023).
- **Sesgo de raza y etnia:** Se evidencia en algoritmos que favorecen a ciertos grupos raciales. (Kotek et al., 2024) Los sistemas de diagnóstico asistido han mostrado menor precisión para pacientes negros que blancos. En modelos de lenguaje, la raza muestra el mayor sesgo de asociación, seguida por género, salud y religión (Bai et al. 2025).
- **Sesgo cultural y de valores:** Los LLM, a menudo exhiben un sesgo cultural con "acento americano". Los modelos reflejan valores de la sociedad donde se entrenaron, generando conflictos al implementarse en otros contextos (Jiao et al., 2023).
- **Sesgo de discapacidad y socioeconómico:** Los modelos pueden reflejan prejuicios hacia personas con discapacidad o estatus socioeconómico bajo, impactando procesos cruciales como decisiones judiciales (Navigli et al., 2023).
- **Sesgo interseccional:** Se refiere a las formas de discriminación que ocurren en la intersección de múltiples ejes de identidad y opresión, basandonos en el concepto de interseccionalidad Crenshaw (1990) y las propuestas de Hill Collins (2017), en las que porfundizaremos más adelante. Como por ejemplo el hecho de ser mujer y negra. Este sesgo es crucial porque el rendimiento de los algoritmos empeora significativamente para grupos que están en los márgenes de varias categorías. Investigaciones demuestran que las incrustaciones de palabras contextualizadas contienen una distribución de sesgos humanos interseccionales (Simó, 2024). Es esencial que la auditoría de los sistemas de IA adopte esta perspectiva, ya que una evaluación no interseccional resulta incompleta.

2.2. FASE 2: SESGOS EN EL DESARROLLO Y MODELADO DEL ALGORITMO

Esta fase se centra en cómo el propio diseño y proceso de entrenamiento del algoritmo introduce o amplifica los sesgos, incluso cuando los datos se consideran aceptables. Este es un punto donde las decisiones de diseño pueden tener consecuencias éticas y sociales directas.

Los sesgos en esta fase resultan de las elecciones técnicas hechas por los desarrolladores.

2.2.1. Priorización de la Precisión sobre la Equidad

Los algoritmos suelen estar optimizados para maximizar la precisión general. Sin embargo, esta búsqueda de precisión puede llevar a los modelos a "capitalizar correlaciones casuales o anomalías estadísticas" presentes en el conjunto de datos. (Chu, et al., 2024). Un modelo puede producir resultados precisos basándose en justificaciones incorrectas (como usar el género como una característica discriminatoria si todos los ejemplos positivos de un *dataset*[3] provienen de hombres), lo que resulta en discriminación. (Chu, et al., 2024).

2.2.2. Errores de Programación y Ponderación

El sesgo puede surgir de errores técnicos o de que el desarrollador pondere injustamente ciertos factores en la toma de decisiones del algoritmo, reflejando sus propios sesgos (conscientes o inconscientes).

2.2.3. Amplificación del Sesgo

El proceso de modelado, especialmente en modelos grandes de lenguaje (LLM), puede amplificar los sesgos inherentes presentes en los datos de entrenamiento (IBM Data and AI Team, 2023). El concepto de recursividad, esencial en el pensamiento complejo, explica cómo los modelos entrenados con datos ya influenciados por el mismo algoritmo pueden generar un ciclo de retroalimentación que exacerba los sesgos.(Bustelo, 2024).

2.2.4. Sesgo de Selección o Asunciones

3 Un conjunto de datos (dataset) es una colección masiva de textos o información estructurada o no estructurada que se utiliza para entrenar y evaluar sistemas de Inteligencia Artificial y modelos de machine learning.

Este sesgo puede provenir de las asunciones hechas por el creador del modelo. En el procesamiento de lenguaje natural[4], los modelos entrenados con grandes corpus de texto aprenden las semánticas y, con ellas, los prejuicios de la sociedad que generó ese lenguaje (Caliskan, como se citó en Kotek, 2023).

2.3. FASE 3: SESGOS EN EL USO, IMPLEMENTACIÓN Y RESULTADOS (POST-ALGORÍTMICOS)

En la fase de implementación, los sesgos se reflejan en los resultados y las consecuencias prácticas del uso del sistema en un entorno real. Estos sesgos se clasifican según el tipo de daño que causan. Los tipos de daños y sus manifestaciones son los siguientes:

2.3.1. Daños Asignativos

Se presentan cuando un sistema automatizado distribuye recursos u oportunidades de manera injusta entre diferentes grupos sociales, como vemos en los siguientes ejemplos.

- **Justicia penal:** Un ejemplo es el algoritmo COMPAS, utilizado para predecir la reincidencia, que mostró un sesgo racial al discriminar a las personas afroamericanas (simó, 2024).
- **Contratación:** Amazon eliminó un algoritmo de selección de personal al descubrir que favorecía a los hombres basándose en el lenguaje de sus currículos (Costanza-Chock, et al., 2022).
- **Finanzas y servicios:** Los algoritmos utilizados para la aprobación de préstamos pueden introducir sesgos invisibles que afectan desproporcionadamente a ciertos grupos.
- **Salud:** Los LLM empleados para gestionar la salud de poblaciones han demostrado sesgo racial, afectando la asignación de recursos (Tao, et al., 2024).

2.3.2. Daños Representacionales

Aparecen cuando un sistema representa a grupos sociales de manera desfavorable, los denigra, perpetúa estereotipos o directamente no reconoce su existencia (Blodgett et al., 2020).

- **Generación de Imágenes:** Se ha encontrado que aplicaciones de IA generativa (como Midjourney) perpetúan el sesgo de género. Al generar imágenes de profesionales especializados, los mayores de edad eran consistentemente representados como hombres, reforzando estereotipos sobre el rol laboral de la mujer (IBM Data and AI Team, 2023).
- **Estereotipos Lingüísticos:** Los modelos de lenguaje pueden crear narrativas que reflejan sesgos persistentes, como el sesgo anti-musulmán (Bai, et al., 2025).
- **Sesgo Estadístico:** Se produce cuando la discriminación grupal se basa en un hecho estadísticamente relevante en el mundo real, pero cuyo uso por el algoritmo resulta en un tratamiento desventajoso hacia un grupo vulnerable (Eticas Research and Consulting, 2021).

2.3.3. Sesgo Cognitivo en la Implementación

El sesgo cognitivo de los humanos que diseñan y utilizan la IA también influye en la implementación. Los desarrolladores pueden preferir conjuntos de datos recopilados de estadounidenses en lugar de tomar muestras de poblaciones globales diversas (IBM Data and AI Team, 2023). Incluso en la interacción post-algorítmica, los equipos humanos pueden incorporar estos sesgos a través de la selección o ponderación de datos.

El análisis del contexto social, económico y cultural en el que opera el algoritmo es esencial para interpretar los resultados y asegurar la deseabilidad del sistema, garantizando que no incurra en formas de discriminación o impacte negativamente en grupos vulnerables (Eticas Research and Consulting, 2021).

2.4. LA NECESIDAD DE UNA AUDITORÍA INTEGRAL

Para enfrentar y mitigar los sesgos en la inteligencia artificial, es necesario adoptar un enfoque que vaya más allá de lo puramente técnico y se alinee con el pensamiento complejo y la metodología de auditoría de sesgos. La auditoría de sesgos se considera una práctica esencial para verificar sistemáticamente el cumplimiento de criterios objetivos y evaluar la confiabilidad de los algoritmos. Esto implica el uso de métricas de equidad grupal para determinar si el algoritmo opera de manera diferente entre grupos protegidos, así como el análisis de sensibilidad para observar cómo las variaciones en los datos afectan los resultados.

Es crucial que esta auditoría sea un proceso continuo y no un evento único, incorporando un monitoreo constante para detectar sesgos emergentes y asegurar que el modelo siga siendo justo y efectivo con el tiempo. Para lograr un desarrollo ético y equitativo de la inteligencia artificial, el enfoque debe ir más allá de los sesgos tradicionales (como raza y género) e incluir la interseccionalidad.

Es fundamental que los desarrolladores y los profesionales del ámbito social colaboren para que los sistemas de inteligencia artificial no solo sean técnicamente sólidos, sino también socialmente responsables y alineados con los principios de justicia y equidad. Aunque hemos identificado que el sesgo se origina y se amplifica en estas etapas, su detección y corrección requieren un marco analítico más sofisticado que el simple análisis técnico, como hemos mencionado anteriormente. A continuación, se presenta la perspectiva interseccional como la herramienta crítica necesaria para este propósito.

4 El Procesamiento del Lenguaje Natural (NLP) (Natural Language Processing) es un área de la Inteligencia Artificial que se enfoca en el desarrollo de modelos computacionales con la capacidad de comprender y generar lenguaje humano. El campo, que ha experimentado un cambio de paradigma con el advenimiento de los Modelos de Lenguaje a Gran Escala (LLM), aborda una amplia variedad de tareas, incluyendo la traducción, la generación de código y el análisis de sentimientos.

3. LA PERSPECTIVA INTERSECCIONAL Y LA MATRIZ DE DOMINACIÓN COMO HERRAMIENTAS DE ANÁLISIS CRÍTICO

Para superar las limitaciones de los análisis de sesgo unidimensionales, que examinan categorías como la raza o el género de manera aislada, es esencial adoptar la interseccionalidad como una herramienta tanto teórica como metodológica. Este enfoque es vital para hacer visibles las experiencias de opresión que son complejas y cualitativamente distintas, las cuales los métodos tradicionales suelen pasar por alto, permitiendo así un diagnóstico más preciso y profundo de los sesgos en la IA.

3.1. FUNDAMENTOS DE LA INTERSECCIONALIDAD

El término interseccionalidad fue introducido por la académica y jurista Kimberlé Crenshaw en 1990 para explicar cómo diversos sistemas de opresión, como el racismo, el sexismo y el clasismo, se entrelazan y generan experiencias de discriminación únicas y complejas.

Crenshaw emplea la metáfora de una encrucijada para mostrar que la discriminación puede surgir desde múltiples direcciones. En este contexto, la interseccionalidad no es simplemente la suma de opresiones (por ejemplo, racismo + sexismo), sino una interacción sinérgica que origina una nueva forma de subordinación. Esto implica que el efecto combinado de varios sesgos no es meramente aditivo (por ejemplo, sesgo (raza) + sesgo (género)), sino multiplicativo o transformador, creando una forma de discriminación cualitativamente distinta y, a menudo, más grave.

Un ejemplo concreto que ilustra esta dinámica, tomado de los estudios de Crenshaw, es el caso de las mujeres inmigrantes que sufren violencia doméstica. Estas mujeres enfrentan una doble subordinación: por un lado, la violencia de género por parte de sus parejas, y por otro, el temor a la deportación debido a su estatus migratorio. Esta situación genera una vulnerabilidad única que no puede ser comprendida si se analiza únicamente desde la perspectiva del género o de la clase social por separado.

3.2. LA INSUFICIENCIA DEL ANÁLISIS DE EJE ÚNICO EN LA INTELIGENCIA ARTIFICIAL

Los enfoques que auditan el sesgo de forma aislada son incapaces de detectar las formas más graves de discriminación algorítmica. El trabajo pionero de Buolamwini y Gebru (2018) es un ejemplo paradigmático. En su estudio sobre clasificadores faciales comerciales, demostraron que estos sistemas obtenían las tasas de error más altas con mujeres de piel oscura. Este fallo garrafal, que afectaba a un grupo en la intersección de raza y género, solo pudo ser detectado gracias a una auditoría interseccional que desagregó los resultados por ambas categorías simultáneamente (Buolamwini y Gebru, 2018).

Investigaciones técnicas posteriores han corroborado estos hallazgos. Por ejemplo, Tan y Celis (2019) demostraron que las identidades interseccionales (cómo ser mujer y afroamericana) sufren sesgos significativamente mayores en las representaciones de palabras contextualizadas generadas por LLM, en comparación con los sesgos asociados a sus identidades constituyentes por separado.

La creciente ubicuidad de la inteligencia artificial en diversos ámbitos de la sociedad, enfatiza la imperiosa necesidad de abordar su desarrollo desde una perspectiva interseccional. Este enfoque implica considerar la interrelación de las diversas identidades y experiencias sociales en la interacción con la IA.

Garantizar un desarrollo de la IA con mayor integridad requiere el reconocimiento y abordaje de las disparidades existentes en términos de género, raza, clase social, discapacidad y otras dimensiones de diversidad. Esto no solo contribuiría a mitigar los sesgos inherentes en los sistemas de IA, sino que también fomentaría soluciones más equitativas y representativas. La adopción de esta visión interseccional en el desarrollo de la IA es esencial para la creación de tecnologías verdaderamente inclusivas y beneficiosas para el conjunto de la sociedad, evitando así la perpetuación o exacerbación de las desigualdades existentes.

3.3. APLICANDO LA INTERSECCIONALIDAD ESTRUCTURAL, POLÍTICA Y REPRESENTACIONAL A LA IA

La aplicación del enfoque interseccional en el desarrollo de la IA requiere analizar cómo las dimensiones de identidad y experiencia social se entrelazan y afectan la interacción con sistemas de IA. La adaptación del marco interseccional al contexto de la IA, implica examinar cómo las estructuras, políticas y representaciones pueden perpetuar o desafiar desigualdades existentes. Esto incluye analizar las prácticas de contratación en la industria, las políticas de desarrollo y despliegue de sistemas, y la representación de grupos en los conjuntos de datos para entrenar modelos de lenguaje.

Para aplicar este marco a los modelos de lenguaje de gran escala (LLM), adaptamos la triple clasificación de Crenshaw (1990) al contexto de la IA.

- **Interseccionalidad estructural:** Se refiere a cómo la estructura de los sistemas de IA, desde bases de datos hasta arquitectura, puede excluir a grupos en las intersecciones. Si los datos no reflejan la realidad de mujeres de color con bajos ingresos, el LLM será incapaz de generar respuestas pertinentes a sus necesidades, perpetuando su invisibilidad estructural.
- **Interseccionalidad política:** Expone cómo las agendas de mitigación de sesgos, al operar sobre un único eje, marginan a quienes están en las intersecciones. La política antirracista tiende a centrarse en hombres negros, mientras la feminista en mujeres blancas, volviendo invisibles a las mujeres negras.
- **Interseccionalidad representacional:** Analiza cómo los LLM generan estereotipos culturales dañinos que afectan a grupos interseccionales. Un modelo que asocia mujeres negras con agresividad o latinas con hipersexualización perpetúa una violencia representacional con consecuencias tangibles. Comprendido el poder analítico de la interseccionalidad, el siguiente paso es integrarla en un marco práctico de gobernanza como la auditoría algorítmica.

4. INTEGRACIÓN DE LA INTERSECCIONALIDAD EN LA AUDITORÍA ALGORÍTMICA

La auditoría algorítmica es un proceso sistemático de evaluación y análisis de los algoritmos utilizados en sistemas de inteligencia artificial y toma de decisiones automatizada. Su objetivo principal es examinar la transparencia, equidad y responsabilidad de

estos algoritmos, identificando posibles sesgos, errores o impactos negativos en diferentes grupos sociales. Este proceso implica la revisión del código fuente, los datos de entrenamiento y los resultados producidos por el algoritmo, así como la evaluación de su conformidad con estándares éticos y regulaciones legales. La auditoría algorítmica es fundamental para garantizar que los sistemas basados en algoritmos sean confiables, justos y no discriminatorios, especialmente en áreas sensibles como la justicia penal, la contratación laboral o la asignación de recursos públicos.

La auditoría algorítmica se posiciona como un mecanismo de gobernanza crucial para la rendición de cuentas de los sistemas de IA. Sin embargo, para que estas auditorías sean verdaderamente efectivas, deben trascender las métricas técnicas estándar y adoptar un análisis sociotécnico robusto. En este contexto, la interseccionalidad no es solo un concepto teórico deseable, sino una categoría de análisis indispensable para evaluar la equidad real de un sistema.

4.1. PRINCIPIOS DE UNA AUDITORÍA ALGORÍTMICA

De acuerdo con la Guía de Auditoría Algorítmica de Éticas (2021), una auditoría exhaustiva debe estar orientada por un conjunto de principios fundamentales que aseguren su integridad y su impacto social positivo. Estos principios incluyen:

- Cumplimiento legal y ético: Garantizar que el sistema algorítmico se adhiera a la normativa vigente y a los estándares éticos aplicables.
- Deseabilidad: Verificar que el algoritmo no incurra en formas de discriminación, especialmente contra individuos o grupos vulnerables.
- Aceptabilidad: Asegurar que el funcionamiento del sistema sea transparente y comprensible para la ciudadanía, permitiendo una supervisión y un control social efectivos.

4.2. UN MARCO INTERSECCIONAL PARA LA AUDITORÍA. PUNTOS DE INTEGRACIÓN

El auge de los sistemas de Inteligencia Artificial para el tratamiento de información ha coincidido con la creciente relevancia empírica del enfoque interseccional en el ámbito social, incitando la búsqueda de un punto de confluencia entre ambos métodos de análisis. La interseccionalidad, ofrece una herramienta analítica esencial para comprender y examinar la complejidad del mundo y de las experiencias humanas, advirtiendo que la desigualdad social está determinada por múltiples ejes (como raza, género o clase) que actúan de manera conjunta y se influencian entre sí (Hill Collins y Bilge, 2019).

Dada la injerencia de los modelos de aprendizaje automático en esferas críticas de la vida social y profesional, resulta pertinente interrogarse acerca de la posible relación simbiótica entre interseccionalidad e IA (Simó, 2024). El enfoque interseccional es crucial para operar la justicia de manera efectiva en la IA (Ovalle et al., 2023), y su ausencia en el proceso auditor resulta en una evaluación incompleta (Costanza-Chock, 2022; Simó, 2024).

La literatura especializada, en consonancia con las directivas de buenas prácticas y los estudios de equidad interseccional, propone que esta perspectiva debe ser tejida en el ciclo completo de la IA (Wang et al., 2023; Ovalle et al., 2023). Específicamente, los estudios sugieren tres momentos clave en el ciclo de vida de un sistema de IA donde la interseccionalidad debe ser evaluada de forma explícita (Simó, 2024), que se describen a continuación:

4.2.1. En la configuración de las bases de datos de entrenamiento

La auditoría debe verificar si los datos de entrenamiento representan de manera adecuada la diversidad de la población, prestando especial atención a los subgrupos interseccionales (Simó, 2024). Esto se alinea con la práctica cuantitativa esencial de los auditores, quienes informan que evalúan la representatividad de los datos y el sesgo en los datos de entrada (Costanza-Chock et al., 2022).

La meta es conseguir una representación fiel del entorno mediante la obtención de datos desagregados (Simó, 2024). Es crucial evitar que los datos de grupos marginalizados sean descartados o malinterpretados como "ruido" estadístico, ya que la heterogeneidad de los datos puede traducirse en una pérdida de precisión en el entrenamiento de los sistemas, disminuyendo el rendimiento para estas categorías (Buolamwini y Gebru, 2018). Esta etapa busca evitar la sobrerrepresentación de categorías hegemónicas y abordar el problema de desequilibrio de clases, frecuente al intentar configurar una base de datos interseccional (Simó, 2024).

4.2.2. En el descubrimiento de correlaciones

Durante la fase de desarrollo y entrenamiento del modelo, se debe investigar activamente si el algoritmo está aprendiendo correlaciones espurias o patrones emergentes que perjudican a subgrupos interseccionales específicos (Simó, 2024).

El potencial de los modelos de *machine learning* para el tratamiento de datos multidimensionales y la estimación de correlaciones no aparentes puede ser aprovechado por la interseccionalidad. Sin embargo, esta fase es donde se manifiesta que los modelos de lenguaje contextualizados contienen una distribución de sesgos humanos interseccionales (Tan y Celis, 2019). Por ello, se han desarrollado pruebas de asociación de *embeddings*[5] dirigidas específicamente a identidades interseccionales (Tan y Celis, 2019), siendo necesario examinar si el algoritmo está "capitalizando correlaciones casuales o anomalías estadísticas" presentes en los datos (Bustelo, 2024).

5 El concepto de embedding (o incrustación de palabras) es una técnica de aprendizaje automático utilizada en el procesamiento de lenguaje natural (NLP) donde las palabras o frases se representan vectorialmente con números reales, lo que permite capturar los matices del lenguaje y sirve como un componente fundamental en los Modelos de Lenguaje a Gran Escala (LLM)

4.2.3. En la fase de auditoría como categoría de fiabilidad

La interseccionalidad debe ser un criterio fundamental para certificar la fiabilidad y equidad de un sistema (Simó, 2024). Las auditorías algorítmicas, que consisten en la verificación sistemática del cumplimiento de criterios objetivos, buscan certificar la confiabilidad de los algoritmos.

Un algoritmo no puede ser considerado fiable si demuestra un rendimiento deficiente o sesgado para cualquier subgrupo interseccional relevante (Simó, 2024). Esto se demuestra empíricamente: al examinar el rendimiento de los algoritmos asociados a diferentes atributos en la intersección de raza y género, se ha comprobado que los peores resultados se obtuvieron consistentemente con mujeres negras, remarcando la importancia de realizar auditorías interseccionales para abordar las disparidades en la precisión (Buolamwini y Gebru, 2018; Simó, 2024).

Por lo tanto, es crucial que la auditoría de sesgo se integre como una práctica continua (Raji et al., 2020), y no se limite a medir la equidad en función de subgrupos aislados, sino que incorpore la perspectiva interseccional como una sensibilidad analítica para garantizar que los sistemas de IA no refuercen las desigualdades existentes (Ovalle et al., 2023). Para conseguir un desarrollo ético y equitativo de la IA, el foco debe ir más allá de los ejes de sesgo tradicionales para incluir la interseccionalidad (Ovalle et al., 2023)

5. PROPUESTA METODOLÓGICA PARA UNA AUDITORÍA INTERSECCIONAL

Integrar la interseccionalidad como marco de análisis es crucial para la operatividad efectiva de la equidad en la Inteligencia Artificial (Ovalle et al., 2023; Simó, 2024). Este enfoque es vital para entender cómo la desigualdad se configura a través de múltiples ejes, como la raza, el género y la clase, que actúan conjuntamente e influyen mutuamente (Hill Collins y Bilge, 2019). La falta de esta perspectiva lleva a una evaluación incompleta de los sistemas algorítmicos (Costanza-Chock, 2022; Simó, 2024). Al aplicar una lente interseccional a las fases de la auditoría propuestas en la Guía de Auditoría Algorítmica (Eticas Research and Consulting, 2021), se obtiene una metodología más robusta y completa, diseñada para asegurar que la IA se desarrolle de manera ética y equitativa (Ovalle et al., 2023; Simó, 2024).

5.1. FASE DE ESTUDIO PRELIMINAR Y MAPEO. EL CONTEXTO SITUADO Y LA INTERSECCIONALIDAD ESTRUCTURAL

La etapa inicial de la auditoría, dedicada a la identificación de grupos vulnerables, debe trascender las categorías amplias y genéricas (como "inmigrantes" o "mujeres") (Eticas Research and Consulting, 2021). Es menester, en cambio, definir y analizar subgrupos interseccionales específicos y relevantes para el contexto de aplicación del algoritmo (Simó, 2024; Eticas Research and Consulting, 2021).

El concepto de interseccionalidad estructural (Crenshaw, 1990) es crucial en esta fase, pues subraya cómo la posición de las personas en la intersección de múltiples sistemas de opresión (como raza y género) hace que sus experiencias de subordinación sean cualitativamente distintas (Crenshaw, 1989; Hill Collins, 2017). Por ejemplo, en un sistema de asignación de ayudas sociales, un subgrupo relevante podría ser "mujeres inmigrantes de bajos ingresos que no hablan el idioma local". Esta definición situada reconoce que factores como el idioma, la clase y el género convergen para estructurar las vivencias de la pobreza de una manera concreta.

La auditoría debe adoptar un enfoque intra categórico, que se centra en los grupos sociales situados en la intersección de distintas categorías para comprender la complejidad con la que estas se interconectan en sus vidas (Martínez-Palacios, 2017). Para ello, es indispensable que los auditores se sitúen en el contexto sociohistórico para identificar los ejes de desigualdad que están actuando con mayor o menor preponderancia.

La literatura sobre auditoría reconoce esta necesidad, pues uno de los principales desafíos es la subrepresentación de grupos interseccionales o raros en los datos (Raji et al., 2020; Costanza-Chock et al., 2022). Los auditores, al definir los grupos a monitorear, deben reconocer que estos pueden ser definidos por la pertenencia a múltiples atributos protegidos (Eticas Research and Consulting, 2021) y deben ser identificados de manera dinámica durante el proceso, en correspondencia con la realidad en la que se inscribe el algoritmo (Eticas Research and Consulting, 2021).

5.2. FASE DE PLAN DE ANÁLISIS Y EJECUCIÓN. DESAGREGACIÓN DE MÉTRICAS DE EQUIDAD

Durante la fase de análisis cuantitativo, el objetivo es identificar el sesgo algorítmico (Eticas Research and Consulting, 2021). Es fundamental desagregar las métricas de rendimiento y equidad, tales como las tasas de falsos positivos (FPR) y falsos negativos (FNR), considerando sus intersecciones y no únicamente ejes individuales como género o raza (Eticas Research and Consulting, 2021).

Este enfoque de equidad interseccional permite identificar disparidades que podrían permanecer ocultas en un análisis unidimensional (Simó, 2024). El análisis debe comparar el desempeño del modelo entre subgrupos demográficos (Bustelo, 2024), revelando si el algoritmo se comporta de manera diferente entre grupos protegidos (Eticas Research and Consulting, 2021).

La investigación ha demostrado que los algoritmos presentan un rendimiento inferior para subgrupos en la intersección (Buolamwini y Gebru, 2018; Ovalle et al., 2023). Los peores resultados en clasificadores de género se observaron en mujeres negras, lo que evidencia la necesidad de auditorías interseccionales (Buolamwini y Gebru, 2018; Simó, 2024). El análisis debe examinar variables proxy[6] que, de manera aislada, carecen de interés, pero que en conjunto revelan información sensible (Eticas

6 En un sistema de inteligencia artificial, una variable proxy es una medición indirecta que se utiliza en lugar de la variable real que se desea optimizar o predecir, generalmente porque esta última es difícil de observar o cuantificar. Por ejemplo, un algoritmo puede usar el nivel de ingresos como variable proxy de la capacidad adquisitiva, o el tiempo de permanencia en una web como proxy del interés del usuario. El uso de variables proxy puede introducir sesgos o distorsiones, ya que no siempre representan fielmente la realidad que se intenta modelar.

Research and Consulting, 2021). El informe debe explicitar qué variables proxy captura el sistema y cómo las combina, evitando que se capitalicen correlaciones casuales (Eticas Research and Consulting, 2021; Bustelo, 2024).

Este enfoque se alinea con la flexibilización de la vigilancia intelectual (Ovalle et al., 2023) mediante el uso de metodologías que preserven las características sociales e históricas de los grupos interseccionales (Ovalle et al., 2023).

Las variables proxy son indicadores que, de manera aislada, pueden carecer de interés, pero que, al ser examinados junto a otras variables, revelan información sensible mediante inferencia (Eticas Research and Consulting, 2021).

5.3. FASE DE INFORME DE AUDITORÍA. COMUNICACIÓN DE LA COMPLEJIDAD

El Informe de auditoría (Eticas Research and Consulting, 2021) debe comunicar de manera explícita y clara los resultados del análisis interseccional. Esto es parte de las responsabilidades de transparencia y rendición de cuentas de los sistemas de IA (Eticas Research and Consulting, 2021).

El informe debe destacar qué variables *proxy* o qué intersecciones de identidades fueron más determinantes para los resultados del sistema, y señalar los riesgos de discriminación asociados a estos hallazgos (Eticas Research and Consulting, 2021). La publicación de la información detallada ayuda a combatir la opacidad algorítmica y permite a los equipos humanos comprender el impacto social de los modelos (Eticas Research and Consulting, 2021).

Además, la interseccionalidad debe ser un criterio fundamental para certificar la fiabilidad y equidad de un sistema (Simó, 2024). Un algoritmo no puede ser considerado fiable si demuestra un rendimiento deficiente o sesgado para cualquier subgrupo interseccional relevante (Simó, 2024).

6. DESAFÍOS PRÁCTICOS Y CONCEPTUALES DE LA AUDITORÍA INTERSECCIONAL

Si bien la integración de la interseccionalidad en la auditoría algorítmica es una estrategia poderosa para promover la equidad (Ovalle et al., 2023), su implementación no está exenta de importantes desafíos prácticos y conceptuales (Simó, 2024).

6.1. REDUCCIONISMO Y SIMPLIFICACIÓN

Existe una tendencia en la investigación de IA a reducir la interseccionalidad a una simple "equidad de subgrupos", lo que se limita a optimizar métricas sobre grupos demográficos sin abordar la realidad social subyacente que impulsa la opresión.

Esta simplificación puede llevar a la despolitización del enfoque, olvidando su raíz genealógica en el feminismo negro y su énfasis en la justicia social (Ovalle et al., 2023).

6.2. PROBLEMAS DE MUESTREO Y DATOS RAROS

Los auditores señalan que los problemas de tamaño de muestra son muy reales para las clases demográficas pequeñas, que a menudo son los grupos interseccionales más vulnerables (Costanza-Chock et al., 2022). Si los datos de estos grupos marginalizados son escasos, pueden ser malinterpretados como "ruido" estadístico (Simó, 2024), lo que dificulta la modelización robusta y la capacidad del algoritmo para generalizar a estas poblaciones (Buolamwini y Gebru, 2018).

6.3. DIFICULTADES LEGALES Y DE RECOLECCIÓN DATOS

Algunos auditores expresan preocupación sobre la posibilidad de infringir la legislación antidiscriminación al intentar recopilar los datos demográficos necesarios para realizar un análisis interseccional detallado, especialmente en categorías que no están legalmente protegidas (Costanza-Chock et al., 2022).

Para mitigar estos desafíos, se requiere que la auditoría no solo se concentre en el análisis técnico, sino que también incorpore el pensamiento complejo para comprender la interconexión y las múltiples dimensiones de los sesgos algorítmicos (Bustelo, 2024). La auditoría debe ser un proceso continuo, incluyendo la revisión constante del impacto del modelo en entornos dinámicos y fomentando la participación de las comunidades que podrían ser directamente perjudicadas por el sistema (Costanza-Chock et al., 2022).

7. DESAFÍOS METODOLÓGICOS EN LA INTEGRACIÓN

La operativización de la interseccionalidad en la auditoría algorítmica se confronta con varios retos metodológicos clave, muchos de los cuales reflejan tensiones epistemológicas históricas entre las ciencias sociales críticas y las ciencias computacionales.

7.1. LA TENSIÓN ENTRE LA TEORÍA CRÍTICA Y LA CUANTIFICACIÓN TÉCNICA

El desafío más acuciante radica en cómo los investigadores de IA, que a menudo operan bajo epistemologías arraigadas en el positivismo y el empirismo (Ovalle et al., 2023), interpretan y operacionalizan un marco conceptual cuyo objetivo fundacional es la justicia social y la praxis (Crenshaw, 1990; Ovalle et al., 2023).

7.1.1. Reducción a la equidad de subgrupos

Cómo incide Ovalle (2023), una revisión crítica de la literatura revela que la mayoría de los investigadores reducen abrumadoramente la interseccionalidad a la optimización de métricas de equidad sobre subgrupos demográficos. Esto conlleva a menudo a la confusión entre la interseccionalidad y la equidad de subgrupos.

7.1.2. Desconexión estructural

Esta reducción a la cuantificación simplifica la crítica, neutralizando el potencial del concepto que fue diseñado para examinar las relaciones de poder que crean y refuerzan las desigualdades. Al operar de esta manera, los investigadores fallan en conectar explícitamente los subgrupos a las estructuras sociales, el contexto o la historia que causan la desigualdad (Ovalle et al., 2023).

7.2. EL PROBLEMA DE LAS CATEGORIZACIONES RÍGIDAS

La interseccionalidad, en su origen, buscó romper la visión monolítica de la mujer, evitando la esencialización de las categorías y la cosificación de las identidades (Simó, 2024). Sin embargo, al aplicarse a la IA, surge el riesgo de la simplificación y neutralización del concepto (Ovalle et al., 2023; Parra y Busquier, 2022).

7.2.1. Neutralización y despolitización

El concepto se ha expandido y masificado de manera acrítica en el ámbito académico y en las políticas públicas, lo que conlleva su simplificación y neutralización (Parra y Busquier, 2022). Se busca una interpretación heurística y vaga, fácil de operacionalizar, lo que neutraliza la capacidad del concepto para confrontar las estructuras de poder.

7.2.2. La Necesidad de flexibilidad

La complejidad intrínseca de la interseccionalidad obliga a rechazar las categorías estáticas y las categorizaciones rígidas. La IA, al reducir las identidades a atributos fijos para su modelado, a menudo diluye la intencionalidad original del marco, que consiste en cuestionar cómo las estructuras sociales influyen en las diferencias grupales percibidas (Ovalle et al., 2023).

7.2.3. Retos de datos raros

Incluso cuando los auditores desean implementar el análisis interseccional (el 65% de los auditores expresan interés) (Costanza-Chock et al., 2022), se enfrentan a desafíos prácticos como la escasez de datos en los subgrupos interseccionales (Simó, 2024; Costanza-Chock et al., 2022). Los datos de grupos marginalizados con baja prevalencia pueden ser malinterpretados como "ruido" estadístico, haciendo que la modelización sea inservible para estas poblaciones (Buolamwini & Gebru, 2018; Simó, 2024; Eticas Research and Consulting, 2021).

8. CONCLUSIONES E IMPLICACIONES FUTURAS

Este análisis revela que los sesgos en los Modelos de Lenguaje a Gran Escala constituyen un problema complejo y multifacético, presente a lo largo de todo su ciclo de vida, desde la recopilación de datos hasta su implementación final. Se ha argumentado que los enfoques unidimensionales para la detección de sesgos son insuficientes y, en ocasiones, contraproducentes. La conclusión principal es que la interseccionalidad no debe ser considerada simplemente como una perspectiva teórica deseable, sino como una herramienta analítica esencial para una auditoría y mitigación de sesgos verdaderamente efectiva, que aspire a la justicia social.

Para lograr una inteligencia artificial más equitativa, es imprescindible adoptar un enfoque interdisciplinario que integre de manera genuina los conocimientos de la sociología, la ética y las ciencias de la computación. La solución a los sesgos algorítmicos no es puramente técnica; requiere un compromiso profundo con valores como la justicia, la transparencia, la explicabilidad y el diseño centrado en el ser humano. Solo a través de la colaboración entre disciplinas y un enfoque crítico y reflexivo podremos construir sistemas de inteligencia artificial que sirvan a toda la humanidad y no solo a sus segmentos más privilegiados.

Las implicaciones de este trabajo señalan varias líneas de investigación futuras que son cruciales para avanzar en este campo:

- **Desarrollo de métricas de equidad explícitamente interseccionales.** Es necesario ir más allá de las métricas existentes y crear nuevas formas de cuantificar el sesgo que capturen las disparidades en las intersecciones de identidades.
- **Creación de corpus de entrenamiento más diversos y representativos a nivel global.** Se deben realizar esfuerzos concertados para recopilar y curar datos que reflejen la diversidad lingüística, cultural y social del mundo, prestando especial atención a las comunidades históricamente marginadas.
- **Investigación sobre metodologías de auditoría cualitativas y mixtas.** Se necesitan nuevos métodos de auditoría que puedan manejar la fluidez, el contexto y la naturaleza relacional de las identidades sociales, complementando los análisis cuantitativos con enfoques cualitativos que capturen las experiencias vividas de las personas afectadas por estos sistemas.

9. BIBLIOGRAFÍA

Adilazuarda, M. F., Mukherjee, S., Lavania, P., Singh, S., Aji, A. F., O'Neill, J., Choudhury, M. (2024). Towards measuring and modeling "culture" in LLMs: A survey. *arXiv preprint arXiv:2403.15412*. https://doi.org/10.48550/arXiv.2403.15412

Bai, X., Wang, A., Song, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.2416228122

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.2005.14050

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

Bustelo, J. L. (2024). La intersección entre la inteligencia artificial (IA), el pensamiento complejo y la metodología de auditoría de sesgo. *Revista Iberoamericana De Complejidad Y Ciencias Económicas, 2(4), 5-16.* https://doi.org/10.48168/ricce.v2n4p5

Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: A taxonomic survey. *ACM SIGKDD Explorations Newsletter*, *26*(1), 34-48. https://doi.org/10.1145/3682112.3682117

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1571-1583). ACM. https://doi.org/10.1145/3531146.353321

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum, 1989(1), 139–167.
Recuperado de https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

Crenshaw, K. (1990). Mapping the margins: Intersectional identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.3917/cdge.039.0051

Devinney, H., Björklund, J., & Björklund, H. (2024). We don't talk about that: case studies on intersectional analysis of social bias in large language models. In *Workshop on Gender Bias in Natural Language Processing (GeBNLP)* (pp. 33-44). Association for Computational Linguistics. https://aclanthology.org/2024.gebnlp-1.3/

Eticas Research and Consulting. (2021). *Guía de Auditoría Algorítmica*. Agencia Española de Protección de Datos. https://eticas.ai/wp-content/uploads/2024/04/Guide-to-Algorithmic-Auditing-SP.pdf

Hill Collins, P., & Bilge, S. (2019). *Interseccionalidad* (ed. en español). Ediciones Morata.

Hill Collins, P. (2017). La diferencia que crea el poder: interseccionalidad y profundización democrática. *Investigaciones Feministas*, *8*(1), 19-39. https://doi.org/10.5209/INFE.54888

Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., & Roozenbeek, J. (2025). Generative language models exhibit social identity biases. *Nature Computational Science*, *5*(1), 65-75. https://doi.org/10.48550/arXiv.2310.15819

IBM Data and AI Team. (2023, 16 de octubre). El sesgo de la IA, a la luz de ejemplos reales [Página Web]. https://www.ibm.com/es-es/think/topics/shedding-light-on-ai-bias-with-real-world-examples

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*. https://doi.org/10.48550/arXiv.2303.17466

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference* (pp. 12-24). ACM. https://doi.org/10.1145/3582269.3615599

Kotek, H., Sun, D. Q., Xiu, Z., Bowler, M., & Klein, C. (2024). Protected group bias and stereotypes in large language models. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2403.14727

Martínez-Palacios, J. (2017). Exclusión, profundización democrática e interseccionalidad. *Revista de Investigaciones Feministas*, *8*(1), 53-71. https://doi.org/10.5209/INFE.54827

Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2023). Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. *arXiv preprint arXiv:2310.13627*. https://doi.org/10.48550/arXiv.2305.14456

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, *15*(2), 1-27. https://doi.org/10.1145/35973

Ovalle, A., Subramonian, A., Gautam, V., Gee, G., & Chang, K.-W. (2023). Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. https://doi.org/10.1145/3600211.3604705

Parra, F., & Busquier, L. (2022). Retrospectivas de la interseccionalidad a partir de la resistencia desde los márgenes. *Las Torres de Lucca. Revista internacional de filosofía política*, *11*(1), 23-35. https://doi.org/10.5209/ltdl.77044

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). *Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing*. En *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 33-44). ACM. https://doi.org/10.1145/3351095.3372873

Rebolledo, J., & Galaz, C. (2022). Interseccionalidad: Aspectos conceptuales y recomendaciones para las políticas públicas. Dirección de Estudios de PRODEMU. https://www.prodemu.cl/wp-content/uploads/2023/05/Interseccionalidad12.pdf

Sakib, S. K., & Das, A. B. (2024). Challenging fairness: A comprehensive exploration of bias in LLM-based recommendations. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 1585-1592). IEEE. https://doi.org/10.1109/BgData62323.2024.10825082

Simó, E. (2024). Hacia una Inteligencia Artificial Interseccional para el tratamiento de información. *Investigaciones Feministas*, *15*(1), 137-144. https://doi.org/10.5209/infe.81954

Tan, Y. C., & Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/201d546992726352471cfea6b0df0a48-Abstract.html

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS nexus*, *3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3730–3748. https://doi.org/10.48550/arXiv.2310.09219

Zhang, D., Zhang, Y., Bihani, G., & Rayz, J. (2024). Hire Me or Not? Examining Language Model's Behavior with Occupation Attributes. *arXiv preprint arXiv:2405.06687*. https://doi.org/10.48550/arXiv.2405.06687

NOTA BIOGRÁFICA

Formador en competencias digitales e Inteligencia Artificial para la intervención social. Profesor del Grado de Trabajo Social en la Universidad Pontificia de Comillas. Graduado en Trabajo Social por la Universidad Complutense de Madrid, Máster en Redes Sociales y Aprendizaje Digital por la UNED, y Máster en Innovación e Inteligencia Artificial por Founderz-Microsoft. Desde 2017 ejerciendo como consultor para organizaciones del Tercer Sector y la administración pública, tras un bagaje de 10 años en el Tercer Sector como Trabajador Social.