

„KI-basierte Systeme verfestigen oftmals menschliche, strukturelle und gesellschaftliche Vorurteile (und skalieren und verstärken diese teilweise sogar). Diese Verzerrungen lassen sich nur schwer reduzieren und können auf individueller, kollektiver und gesellschaftlicher Ebene Schaden anrichten.“

Quellen von Bias in der KI

- Bias in den Daten
- Bias bei der Algorithmenauswahl
- Bias bei der Implementierung

„Fast 9 von 10 Männern und Frauen hegen grundlegende Vorurteile gegenüber Frauen“, und diese Stereotypen spiegeln sich in KI-Systemen wider.

Wie KI Verzerrungen verstärkt

Bei der Arbeitsplatzverdrängung

- Unverhältnismäßig hohe Arbeitslosigkeit in Sektoren, in denen Frauen einen wichtigen Teil der Arbeitskraft bilden

Im KI-Entwicklungsprozess

- Unterrepräsentation von Frauen
- Schwache Regulierung erleichtert den Missbrauch von KI



Bei Entscheidungsprozessen

- Personalauswahl
- Finanz- und Versicherungswesen

Gesellschaftliche Auswirkungen

- Schwächung des sozialen Zusammenhalts
- Geschlechterbezogene Gewalt
- Homogenisierung vulnerabler Gruppen

Wie lassen sich Verzerrungen in Large Language Models (LLM) erkennen?



- Durch die Analyse davon, wie ein LLM unterschiedliche Konzepte miteinander verknüpft
- Durch die Beobachtung, wie ein LLM Texte zu bestimmten Themen generiert



Vorteile von Open-Source-Modellen

- Transparenz und Überprüfbarkeit
- Erleichtert unabhängige Forschung und Benchmarking
- Durch die Community getragene Weiterentwicklung

„Die zunehmende Komplexität von Systemen der künstlichen Intelligenz erfordert rigorosere Anstrengungen, um Fairness in KI-gestützten Entscheidungen und Interaktionen zu erreichen.“

„Es ist entscheidend, bereits in den frühen Phasen des KI-Entwicklungszyklus Maßnahmen zur Verhinderung von Verzerrungen zu ergreifen“

Zielgruppen: Bibliothekspersonal, Lehrpersonal, Studierende

QUELLEN

UNESCO (2024). *Challenging systematic prejudices: an investigation into bias against women and girls in large language models*. <https://unesdoc.unesco.org/ark:/48223/pf0000388971>

Zum GEDIS-Projekt

Diese Ressource ist Teil des europäischen Projekts GEDIS (Gender Diversity in Information Science), das durch die Veröffentlichung von offenen Bildungsressourcen (OER) geschlechterbezogenen Ungleichheiten im Hochschulbereich adressiert. Der Schwerpunkt liegt den Informations- und Bibliothekswissenschaften.

GEDIS - Gender Diversity in Information Science: Challenges in Higher Education. <https://ub.edu/gedis>

Zitation: Villarroja, Anna, Juan-José Boté-Vericad, Maddalena Fedele, Mari Váñez. 2026. *Biaixos de gènere en els models de llenguatge externs*. DOI: [10.5281/zenodo.18171512](https://doi.org/10.5281/zenodo.18171512). *Geschlechterbias in Large Language Models*. Übersetzt von Stefan Dreisiebner, Thomas Mandl, Merle Bruske und Stefanie Ratzesberger. DOI: [10.5281/zenodo.19032475](https://doi.org/10.5281/zenodo.19032475).

Mitfinanziert durch die Europäische Union. Die in dieser OER geäußerten Meinungen und Ansichten sind ausschließlich die der Autor:innen. Sie spiegeln nicht unbedingt die Meinungen und Ansichten der Europäischen Union oder des Spanischen Dienstes für die Internationalisierung des Bildungswesens (SEPIE) wider. Weder die EU, noch die Bewilligungsbehörde können für diese haftbar gemacht werden.



Kofinanziert von der Europäischen Union



GESCHLECHTERBIAS IN LARGE LANGUAGE MODELS

Vereinfachte Lehrnotiz

Ziel der Ressource

Diese Infografik soll das Bewusstsein von Bibliothekspersonal, Lehrpersonal und Studierenden für das Vorhandensein von Geschlechterbias in Systemen der künstlichen Intelligenz (KI) schärfen, insbesondere in Large Language Models (LLM).

Angestrebte Lernergebnisse

- Studierende verstehen, was Bias in KI-Systemen ist und wie er aus Daten, algorithmischen Entscheidungen und Einsatzkontexten entsteht.
- Sie identifizieren konkrete Bereiche, in denen KI Geschlechterungleichheiten reproduziert oder verstärkt.
- Sie analysieren kritisch die Rolle von LLMs bei der Reproduktion von Geschlechterstereotypen.
- Sie erkennen die Bedeutung der Einbeziehung präventiver Massnahmen in den frühen Phasen der Entwicklung von KI-Systemen.

Reflexionsfragen

- Wie wirkt sich ein Mangel an Vielfalt (insbesondere an Geschlechtervielfalt) in KI-Entwicklungsteams auf die Ergebnisse und Entscheidungen von KI-Systemen aus?
- Welche sozialen Risiken birgt der breite Einsatz verzerrungsbehafteter LLMs in Kontexten wie Bildung oder öffentlicher Kommunikation?
- In welchem Ausmaß kann KI sowohl zur Verhinderung als auch zur Förderung von Formen geschlechterbezogener Gewalt beitragen, insbesondere in digitalen Umgebungen?
- Welche regulatorischen, ethischen oder technischen Strategien wären notwendig, um Bias in KI vor dem Einsatz zu reduzieren?
- Welche Unterschiede lassen sich in der Art beobachten, wie KI-Systeme Geschlecht in Abhängigkeit vom kulturellen oder geografischen Kontext darstellen?

Einsatz in der Lehre

Dieses Lehrmaterial zu Geschlechterbias kann als Einstiegsaktivität in Schulungsseminaren für Studierende oder als Teil von Weiterbildungs- und Professionalisierungsangeboten für Bibliothekspersonal eingesetzt werden.

