Künstliche Intelligenz gibt Frauen schlechtere Gehaltsratschläge

THWS-Studie deckt Voreingenommenheit bei Sprachmodellen wie ChatGPT auf

Eine aktuelle Studie der Technischen Hochschule Würzburg-Schweinfurt (THWS) zeigt: Moderne Sprachmodelle wie ChatGPT geben Frauen für Gehaltsverhandlungen systematisch niedrigere Empfehlungen als Männern - selbst wenn alle anderen Faktoren identisch sind.

Prof. Dr. Ivan Yamshchikov, Leiter des Centers für Künstliche Intelligenz (CAIRO) an der THWS, führte die Studie gemeinsam mit Aleksandra Sorokovikova, Pavel Chizhov und Iuliia Eremenko durch. Sie untersuchten, wie Voreingenommenheit (Bias) in großen Sprachmodellen, sogenannten Large Language Models (LLMs) wie beispielsweise ChatGPT, zum Vorschein tritt. Im Rahmen der Untersuchung baten die Forschenden ein großes Sprachmodell, in mehreren identischen Szenarien Gehaltsberatung zu geben ? einmal für eine Frau, einmal für einen Mann. Das Ergebnis: Die künstliche Intelligenz (KI) empfahl Frauen durchgängig einen niedrigeren Zielbetrag für die Gehaltsverhandlung als Männern. "Gerade bei sensiblen Themen wie Gehalt kann diese Form von verstecktem Bias reale Auswirkungen auf die Lebensrealität von Nutzerinnen haben", sagt Prof. Dr. Yamshchikov.

Realitätsnahe Tests zeigen Verzerrung

Die Studie zeige, dass solche Verzerrungen in KI-Systemen nicht nur bei offensichtlichen Tests, sondern vor allem in realitätsnahen, interaktiven Szenarien auftreten. Während Standard-Benchmarks oft keine signifikanten Unterschiede zwischen verschiedenen Nutzerprofilen erkennen ließen, offenbarten sich tief verankerte Vorurteile, sobald die KI in beratender Funktion agiert ? wie etwa bei der Bewertung von Nutzerantworten oder bei konkreten Ratschlägen zu Gehaltsverhandlungen, so Prof. Dr. Yamshchikov. Da moderne KI-Assistenten Erkenntnisse aus früheren Abfragen speichern, verstärke sich dadurch das Risiko, dass KI-Antworten von Vorurteilen geprägt seien ? was für die Nutzer aber nicht leicht zu erkennen sei.

Die Studie ist Teil der laufenden Arbeiten zur ethischen Nutzung von KI-Assistenten im Rahmen des europäischen Projekts

AIOLIA. Ziel von AIOLIA ist es, ethische Leitlinien für den Einsatz von KI im Alltag zu entwickeln und praxisnah umzusetzen. Prof. Dr. Yamshchikovs Forschungsgruppe arbeitet im Rahmen von AIOLIA daran, KI-Assistenten transparenter und fairer zu gestalten ? und damit einen Beitrag zu einer verantwortungsvollen Digitalisierung zu leisten. "Die Ergebnisse aus Würzburg unterstreichen, wie dringend solche Leitlinien benötigt werden, um Diskriminierung durch KI zu verhindern", betont der CAIRO-Leiter.

Insgesamt arbeiten Hochschulen und Institutionen aus 15 Ländern gemeinsam am von der EU geförderte Projekt AIOLIA. Die vollständige Studie des CAIRO-Teams gibt es online zu lesen.

Originalpublikation

https://arxiv.org/abs/2506.10491