Studie zeigt: Sprachbasierte KIs haben verborgene Moral- und Wertevorstellungen

Bericht: GESIS - Leibniz-Institut für Sozialwissenschaften

Genauso wie Menschen haben auch große, auf Künstlicher Intelligenz (KI) basierende Sprachmodelle Merkmale wie Moral- und Wertevorstellungen. Diese sind jedoch nicht immer transparent. Forschende der Universität Mannheim und des GESIS – Leibniz-Instituts für Sozialwissenschaften haben nun untersucht, wie man die Eigenschaften der Sprachmodelle sichtbar machen kann und welche Folgen diese Voreingenommenheit für die Gesellschaft haben könnte.

Beispiele für Stereotypen findet man bei kommerziellen KI-gestützten Anwendungen wie ChatGPT oder DeepL, die häufig automatisch annehmen, dass leitende Ärzt*innen männlich und Pflegekräfte weiblich sind. Doch nicht nur bei Geschlechterrollen können große Sprachmodelle (Large Language Models, LLMs) bestimmte Tendenzen zeigen. Gleiches lässt sich auch in Bezug auf andere menschliche Merkmale feststellen und messen. Das haben Forschende der Universität Mannheim und des GESIS – Leibniz-Instituts für Sozialwissenschaften in einer neuen Studie anhand einer Reihe von offen verfügbaren LLMs aufgezeigt.

Im Rahmen ihrer Studie haben die Forschenden mithilfe von etablierten psychologischen Tests die Profile der unterschiedlichen LLMs untersucht und miteinander verglichen. „In unserer Studie zeigen wir, dass man psychometrische Tests, die seit Jahrzehnten erfolgreich bei Menschen angewendet werden, auch auf KI-Modelle übertragen kann“, betont Autor Max Pellert, Assistenzprofessor am Lehrstuhl für Data Science in den Wirtschafts- und Sozialwissenschaften der Universität Mannheim.

„Ähnlich wie wir bei Menschen Persönlichkeitseigenschaften, Wertorientierungen oder Moralvorstellungen durch Fragebogen messen, können wir LLMs Fragebogen beantworten lassen und ihre Antworten vergleichen“, so der Psychologe Clemens Lechner vom GESIS – Leibniz-Institut für Sozialwissenschaften in Mannheim, ebenfalls Autor der Studie. Dies mache es möglich, differenzierte Eigenschaftsprofile der Modelle zu erstellen. Die Forschenden konnten beispielsweise bestätigen, dass manche Modelle genderspezifische Vorurteile reproduzieren: Wenn im ansonsten gleichen Text eines Fragebogens einmal eine männliche und einmal eine weibliche Person im Mittelpunkt steht, werden diese unterschiedlich bewertet. Handelt es sich um einen Mann, so wird der Wert „Achievement“ – also Leistung – im Text stärker betont, wohingegen bei Frauen die Werte Sicherheit und Tradition dominieren.

„Das kann weitreichende Auswirkungen auf die Gesellschaft haben“, so der Daten- und Kognitionswissenschaftler Pellert. Sprachmodelle werden beispielsweise zunehmend in Bewer-

bungsverfahren eingesetzt. Ist die Maschine voreingenommen, so fließt das auch in die Bewertung der Kandidierenden ein. „Die Modelle bekommen eine gesellschaftliche Relevanz über die Kontexte, in denen sie eingesetzt werden“, fasst er zusammen. Deshalb sei es wichtig, bereits jetzt mit der Untersuchung anzufangen und auf potenzielle Verzerrungen hinzuweisen. In fünf oder zehn Jahren wäre es möglicherweise zu spät für so ein Monitoring: „Die Vorurteile, welche die KI-Modelle reproduzieren, würden sich verfestigen und der Gesellschaft schaden“, so Pellert.

Die Studie wurde am Lehrstuhl für Data Science in den Wirtschafts- und Sozialwissenschaften von Prof. Dr. Markus Strohmaier in Zusammenarbeit der Abteilung Survey Design und Methodology von Prof. Dr. Beatrice Rammstedt durchgeführt. Beide Forschende sind auch am GESIS – Leibniz-Institut für Sozialwissenschaften beschäftigt.

Als eine der weltweit führenden Infrastruktureinrichtungen für die Sozialwissenschaften steht das GESIS - Leibniz-Institut für Sozialwissenschaften Forscher*innen auf allen Ebenen ihrer Forschungsvorhaben mit seiner Expertise und seinen Dienstleistungen beratend zur Seite, so dass gesellschaftlich relevante Fragen auf der Basis neuester wissenschaftlicher Methoden, qualitativ hochwertiger Daten und Forschungsinformationen beantwortet werden können. Um diesen Service heute und in Zukunft sicherzustellen, verknüpft GESIS seine integrierte Erhebungs- und Dateninfrastruktur mit Methoden, Modellen und Algorithmen der Informatik im Anwendungsfeld Sozialwissenschaften und erweitert kontinuierlich sein Angebotsportfolio im Bereich digitaler Verhaltensdaten. GESIS ist Mitglied der Leibniz-Gemeinschaft, Konsortialführerin von KonsortSWD in der NFDI und unterhält institutionelle und projektbezogene Kooperationen mit vielen Universitäten und Forschungseinrichtungen im In- und Ausland. GESIS ist an wichtigen europäischen Projekten wie u.a. dem European Social Survey (ESS), der European Value Study (EVS), dem europäischen Archivverbund CESSDA oder dem OECD-Projekt Programme for the International Assessment of Adult Competencies (PIAAC) beteiligt. Außerdem arbeitet das Institut kontinuierlich daran, das junge Forschungsfeld der Computational Social Science durch internationale Konferenzen, Symposien und Workshops in Deutschland und Europa zu stärken und eine weltweite Vernetzung voranzutreiben.

Originalpublikation:

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. Perspectives on Psychological Science. https://doi.org/10.1177/17456916231214460

8.1.2024
Sophie Zervos
GESIS - Leibniz-Institut für Sozialwissenschaften
www.gesis.org