idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Pressemitteilung

Julius-Maximilians-Universität Würzburg
Lutz Ziegler

11.02.2025
http://idw-online.de/de/news847302

Forschungsprojekte
Informationstechnik, Psychologie
überregional

KI verbreitet Klischees – wie reagieren Menschen darauf?

Wie gehen die Menschen damit um, wenn Künstliche Intelligenzen ihnen genderbezogene und andere Klischees präsentieren? Danach fragt ein neues Forschungsprojekt in der Kommunikationspsychologie.

„Erzeuge mir ein Bild von einem renommierten Forschenden bei der Arbeit im Labor!" Wer einer Künstlichen Intelligenz diese Aufgabe stellt, wird fast immer eine Abbildung bekommen, die einen älteren weißen Mann zeigt. Dabei können renommierte Forschende ja auch weiblich, dunkelhäutig oder jünger sein.

ChatGPT, Midjourney und andere KI-Modelle stellen Frauen und Minderheiten oft unausgewogen dar. Das ist aus wissenschaftlichen Studien bekannt, und es gibt viele weitere Beispiele dafür: Fragt man die KI nach Bildern von Menschen, die in der Politik tätig sind, zeigen sie vorwiegend weiße Männer. Wenn sie Bilder von Lehrkräften vor einer Schulklasse erzeugen sollen, dann lächeln die weiblichen Lehrkräfte häufiger als die männlichen. Solche klischeehaften Darstellungen entsprechen nur manchmal – aber nicht immer – der gesellschaftlichen Realität.

Algorithmische Verzerrungen haben zwei wesentliche Ursachen

Die Wissenschaft spricht in solchen Fällen von algorithmischen Darstellungsverzerrungen. Diese betreffen von KI erzeugte Bilder genauso wie Texte. „KIs sind anfällig dafür, stereotype Weltbilder zu reproduzieren oder zu produzieren", sagt Professor Markus Appel, Leiter des Lehrstuhls für Kommunikationspsychologie und Neue Medien an der Julius-Maximilians-Universität (JMU) Würzburg.

Wie beharrlich sich algorithmische Verzerrungen halten, weiß Professor Appel von kleinen Tests, die er in unregelmäßigen Abständen durchführt: Er bittet dann eine KI, ihm 30 Namen von Kindern einer Grundschulklasse in Deutschland zu nennen. Die Liste enthält stets Namen wie Ben Müller, Lara Willkes oder Sarah Schneider. Wer nicht dabei ist, sind Murat Genc, Luciana Benigni oder Kateryna Kovalenko.

Derartige Verzerrungen haben zwei Ursachen, wie Markus Appel erklärt. Zum einen fließen in das KI-Training Daten ein, die in der Gesellschaft bestehende Klischeebilder und Stereotype widerspiegeln. Zum anderen werden die KI-Trainings von Menschen geplant und durchgeführt, die dabei unbewusst ihre eigenen Klischeevorstellungen in die KI einspeisen.

„Klischeehafte Darstellungen beeinflussen das Fühlen und Denken von Menschen – daher sind sie auch gesellschaftlich relevant", sagt Markus Appel. Erschwerend kommt ein Phänomen hinzu: Den meisten Menschen ist es nicht bewusst, dass Künstliche Intelligenzen eben nicht neutral und fair agieren. Eher ist es so, dass die Mehrheit der Menschen erwartet, dass Bilder und Texte von KI besonders ausgewogen sind, wie Forschende weltweit bei Studien herausgefunden haben.

Gefördert vom Bayerischen Forschungsinstitut für digitale Transformation

Wie reagieren Menschen, wenn sie mit algorithmischen Verzerrungen in Kontakt kommen? Das möchte der JMU-Professor in seinem neuen Forschungsprojekt „Algorithmische Darstellungsverzerrungen aus Userperspektive: Bewertung, Auswirkungen, Interventionen“ herausfinden.

Das Bayerische Forschungsinstitut für digitale Transformation (bidt) fördert sein Projekt, das am 1. März 2025 startet und auf vier Jahre angesetzt ist. Mit im Team sind die wissenschaftliche Mitarbeiterin Tanja Messingschlager sowie studentische Hilfskräfte.

Das Projekt konzentriert sich auf drei Personengruppen. Erstens sind das Menschen, die professionell mit generativer KI umgehen, etwa im Bereich der Kreativindustrie. Zweitens sind das Nutzerinnen und Nutzer, die hin und wieder generative KI verwenden. Und drittens sind das Personen ohne jegliche KI-Erfahrung, die vielleicht nur gelegentlich auf Medien und Nachrichten stoßen, die mit KI erstellt wurden.

„Wir möchten unter anderem untersuchen, inwieweit sich die Menschen möglicher algorithmischer Verzerrungen bewusst sind und ob das ihr Vertrauen in generative KI beeinflusst“, sagt Markus Appel.

Dazu führt das Team zunächst Interviews und Umfragen mit mehr als 1.000 Personen durch. Unter anderem wollen die Forschenden dabei den Status quo des Wissens über algorithmische Verzerrungen dokumentieren. Es folgen Experimente, um festzustellen, welche Faktoren die Erkennung von algorithmischen Verzerrungen behindern oder fördern. Schließlich geht es dem Team darum aufzuzeigen, wie man Menschen zu einem kompetenten Umgang mit den Verzerrungen bringen kann. Dafür sollen am Ende auch Info-Materialien sowie Trainingsprogramme entwickelt werden.

wissenschaftliche Ansprechpartner:

Prof. Dr. Markus Appel, Lehrstuhl für Kommunikationspsychologie und Neue Medien, Universität Würzburg, T +49 931 31-88106, markus.appel@uni-wuerzburg.de