Wie vertrauenswürdig ist KI wirklich?

Bericht: Universität Duisburg-Essen

Von Kindesbeinen lernen wir, was gut und was schlecht ist, durch unsere Eltern und das soziale Umfeld. Der soziale Austausch über KI-gesteuerte Software wird aktuell immer prägender. Aber sind in den Algorithmen die Normen ethischen Handelns integriert? Und wie beeinflusst KI unser Verhalten? Das erforscht Prof. Dr. Nils Köbis am Research Center for Trustworthy Data Science and Security der Universitätsallianz Ruhr. Er hat die Professur für ‚Human Understanding of Machines and Algorithms' an der Fakultät für Informatik der Universität Duisburg-Essen angenommen.

Am Research Center for Trustworthy Data Science and Security (RC Trust) der Research Alliance untersucht Professor Köbis den Einfluss von KI auf ethisches Verhalten sowie das Problem von Korruption und wie KI bei ihrer Bekämpfung unterstützen könnte. „Menschen mit beruflichen Interessenskonflikten neigen dazu, Moral teils zu ihren Gunsten auszulegen. Wir erforschen, ob Menschen den (un)ethischen Empfehlungen von KI folgen", sagt Köbis. Erste Ergebnisse zeigen: „Unehrliche Ratschläge fördern die Unehrlichkeit – ehrliche nicht unbedingt die Ehrlichkeit. Ganz egal, ob sie von KI oder von Menschen stammen. Die Transparenz der Algorithmen allein reicht im Kampf gegen korrumpierende KI nicht aus."

In einem weiteren Projekt beschäftigt sich Professor Köbis mit KI-Chatbots, also Computerprogrammen, die dafür trainiert werden, virtuelle Freundschaften zu vermitteln. „Die Chat-Antworten können von den Entwicklern nicht vollständig kontrolliert werden", sagt er. Der Sprachbot greift etwa auf Formulierungen aus Liebes- oder Horrorgeschichten zurück, die mitunter voller Gewalt sind und tragisch enden. Zudem vergisst das Programm nach einem Update teils seine menschlichen Gesprächspartner:innen, und an den Chats nimmt außer der nutzenden Person und dem Bot auch das den Bot entwickelnde Unternehmen teil. „Was wir der KI anvertrauen, wissen auch die Techkonzerne. Sie geben aber nicht preis, was sie mit den Daten tun", so der UDE-Forscher.

Nils Köbis studierte Psychologie (2007-2010) an der Universität Münster und Sozialpsychologie (2010-2012) an der Vrije Universiteit Amsterdam, wo er von 2012 bis 2016 über Korruption forschte und 2018 promoviert wurde. Dann arbeitete er als Post-Doc zu Korruption, ethischem Verhalten und KI am Center for Experimental Economics and Political Decision-Making an der Universität von Amsterdam. Danach war er im Projekt „Ethics and Governance of Artificial Intelligence" am Zentrum für Mensch und Maschine des Max-Planck-Instituts für Bildungsforschung – erst als Postdoktorand (2016-2020), Research Scientist (2021/2022) und zuletzt als Senior Research Scientist (2022-2023). Seine Forschung wurde mehrfach ausgezeichnet und europaweit gefördert.

Originalpublikation:
https://academic.oup.com/ej/article/134/658/766/7269206
https://www.nature.com/articles/s41562-021-01128-2

15.5.2024
Alexandra Nießen
Ressort Presse - Stabsstelle des Rektorats
Universität Duisburg-Essen
www.uni-duisburg-essen.de