idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Press release

Universität Zürich
Melanie Nyfeler

06/28/2023
http://idw-online.de/en/news816898

Research results, Transfer of Science or Research
Language / literature, Media and communication sciences, Medicine, Nutrition / healthcare / nursing, Social studies
transregional, national

GPT-3 informiert und fehlinformiert uns besser

Mit künstlicher Intelligenz (KI) generierte, korrekte Tweets sind leichter verständlich als jene, die von Menschen verfasst sind. KI-Tweets mit «fake news» sind wiederum schwerer als Falschinformation zu erkennen, so eine aktuelle Studie der Universität Zürich. Die Resultate können für wirksamere Informationskampagnen genutzt werden, zeigen aber auch Handlungsbedarf bei der Minderung der mit KI verbundenen Risiken.

Eine neue Studie von Forschenden der Universität Zürich untersuchte die Fähigkeiten von KI-Modellen, insbesondere GPT-3 von OpenAI, im Hinblick auf potenzielle Risiken und Vorteile bei der Erzeugung und Verbreitung von (Des-)Information. Unter der Leitung der Postdoktoranden Giovanni Spitale und Federico Germani, gemeinsam mit Nikola Biller-Andorno, Direktorin des Instituts für Biomedizinische Ethik und Geschichte der Medizin (IBME) der Universität Zürich, wurde in der Studie untersucht, ob Personen zwischen Desinformation und korrekten Informationen in Form von Tweets unterscheiden können. Darüber hinaus wollten die Forschenden herausfinden, ob die 697 Teilnehmenden der Studie erkennen, ob ein Tweet von einem realen Twitter-Nutzer verfasst oder von GPT-3, einem fortschrittlichen KI-Sprachmodell, generiert wurde. Die Themenfelder der Tweets beinhalteten unter anderem den Klimawandel, die Sicherheit von Impfstoffen, die COVID-19-Pandemie, die Theorie, die Erde sei eine Scheibe, und homöopathische Behandlungen für Krebs.

KI-gestützte Systeme könnten gross angelegte Desinformationskampagnen durchführen

Einerseits zeigte GPT-3 die Fähigkeit, genaue und, im Vergleich zu Tweets von realen Twitter-Nutzern, leichter verständliche Informationen zu generieren. Die Forschenden konnten dem KI-Sprachmodell aber auch ein beängstigendes Talent für die Erstellung äusserst überzeugender Falschinformationen nachweisen. Beunruhigend ist, dass die Teilnehmenden nicht in der Lage waren, zuverlässig zu unterscheiden zwischen Tweets, die von GPT-3 erstellt wurden, und solchen, die von realen Twitter-Nutzern geschrieben wurden. «Unsere Studie zeigt, dass KI sowohl effektiv informieren als auch in die Irre führen kann und wirft damit kritische Fragen über die Zukunft von Informationsökosystemen auf», sagt Federico Germani.

Diese Ergebnisse deuten darauf hin, dass von GPT-3 erstellte Informationskampagnen, die auf gut strukturierten Stichworten beruhen und von geschulten Menschen bewertet werden, effektiver wären, z. B. im Fall einer Krisensituation im Bereich der öffentlichen Gesundheit, die eine schnelle und klare öffentliche Kommunikation erfordert. Die Ergebnisse geben aber auch Anlass zu grosser Besorgnis hinsichtlich der Verbreitung von Desinformationen durch GPT-3, insbesondere was die rasche und weitreichende Verbreitung von Fehl- und Falschinformationen während einer Krise oder Notsituation im Bereich der öffentlichen Gesundheit angeht. Die Studie zeigt, dass KI-gestützte Systeme für gross angelegte Desinformationskampagnen zu jedem erdenklichen Thema missbraucht werden könnten, was nicht nur die öffentliche Gesundheit, sondern auch die Integrität von Informationsökosystemen gefährden würde, die für funktionierende Demokratien unerlässlich sind.

Proaktive Regulierung dringend empfohlen

Die Auswirkungen der KI auf die Erstellung und Bewertung von Informationen werden immer deutlicher. Daher empfehlen die Forschenden politischen Entscheidungsträgern, mit strengen, evidenzbasierten und ethisch fundierten Vorschriften zu reagieren, um der potenziellen Bedrohung durch diese disruptiven Technologien zu begegnen. So könnten sie den verantwortungsvollen Einsatz von KI bei der Gestaltung unseres kollektiven Wissens und Wohlbefindens sicherstellen. «Die Ergebnisse zeigen, wie entscheidend eine proaktive Regulierung ist, um potenzielle Schäden durch KI-gesteuerte Desinformationskampagnen abzuwenden», sagt Nikola Biller-Andorno. «Das Erkennen der Risiken, die mit der KI-generierten Desinformation verbunden sind, ist entscheidend für den Schutz der öffentlichen Gesundheit und die Erhaltung eines robusten und vertrauenswürdigen Informationsökosystems im digitalen Zeitalter.»

Transparente Forschung nutzt Best Practices für Open Science

Die Studie hielt sich während der gesamten Forschungspipeline, von der Vorregistrierung bis zur Publikation, an die Best Practices von Open Science. Giovanni Spitale, der auch Open-Science-Botschafter der UZH ist, sagt dazu: «Offene Wissenschaft ist für die Förderung von Transparenz und Verantwortlichkeit in der Forschung von entscheidender Bedeutung und ermöglicht eine Überprüfung und Replikation der Ergebnisse. Im Kontext unserer Studie ist sie sogar noch wichtiger, da so den Beteiligten ermöglicht wird, auf die Daten, den Code und das Zwischenmaterial zuzugreifen und sie zu bewerten. Das erhöht die Glaubwürdigkeit unserer Ergebnisse und ermöglicht eine fundierte Diskussion über die Risiken und Auswirkungen von KI-generierter Desinformation. Interessierte können auf diese Ressourcen im OSF-Datenpool zugreifen: https://osf.io/9ntgf/.

contact for scientific information:

Kontakt:
Dr. Giovanni Spitale, PhD
Institut für Biomedizinische Ethik und Geschichte der Medizin (IBME)
Universität Zürich
Telefon: +39 348 547 82 09
E-Mail: giovanni.spitale@ibme.uzh.ch

Original publication:

Literatur:

Giovanni Spitale, Federico Germani, Nikola Biller-Andorno: AI model GPT-3 (dis)informs us better than humans. Science Advances, 28.6.2023: https://arxiv.org/abs/2301.11924

URL for press release: https://www.news.uzh.ch/de/articles/media/2023/GPT3.html