technische universität
dortmund

Medieninformation

Referat Hochschulkommunikation

Baroper Str. 285
D-44227 Dortmund
www.presse.tu-dortmund.de

2025-009 vom 18.02.2025

Vor der Bundestagswahl 2025

Team der TU Dortmund ermittelt Fehleranfälligkeit von KI-Wahlhilfen

Wahlhilfe-Tools, die auf Künstlicher Intelligenz basieren, bieten neue Möglichkeiten, sich interaktiv über die Wahlprogramme der Parteien zu informieren. Doch nach Untersuchungen Forschender der TU Dortmund sind die Antworten der KI nicht rein durch die Inhalte der Parteiprogramme festgelegt. Die KI-Antworten weichen manchmal stark von den Parteiprogrammen ab und sind sehr abhängig von den Eingaben der Nutzenden. Das neunköpfige Team hat getestet, wie zuverlässig zwei KI-Tools die Positionen der Parteien aus dem Wahl-O-Mat wiedergeben: Bei wahl.chat fanden sie in jedem vierten Fall Abweichungen, bei wahlweise.info sogar in jedem zweiten.

Es sind nur noch wenige Tage bis zur Bundestagswahl 2025 und viele Menschen sind noch unsicher, welche Partei sie wählen wollen. Unterstützung bei der Entscheidung bietet seit vielen Jahren der Wahl-O-Mat der Bundeszentrale für politische Bildung. Nutzende entscheiden hier bei 38 Thesen, ob sie diesen zustimmen oder nicht, um zu erfahren, mit welchen Parteien sie am meisten übereinstimmen. In diesem Jahr gibt es daneben nun zwei Angebote, die auf Künstliche Intelligenz setzen: wahlweise.info und wahl.chat. Diese Chat-Programme basieren auf Großen Sprachmodellen wie ChatGPT oder Llama. Die Anwendungen erlauben den Nutzenden völlig freie Fragen zu stellen, um sich über die Wahlprogramme der Parteien und ihre Positionen zu informieren.

Das freie Formulieren von Fragen in den KI-Assistenten bringt indes nach Untersuchungen von Wissenschaftler*innen der TU Dortmund und des Research Center Trustworthy Data Science and Security (RC-Trust) der Universitätsallianz Ruhr Probleme mit sich, die die Vertrauenswürdigkeit der Tools infrage stellen. Sie fanden eine hohe Fehleranfälligkeit bei der Interpretation von Informationen, widersprüchliche Antworten bei der Wiederholung von Fragen oder Lücken im Schutz vor Manipulationsversuchen. „Unsere Sorge ist, dass Bürgerinnen und Bürger den KI-Assistenten vertrauen, obwohl die KI nur die wahrscheinlichste, aber nicht immer die faktisch richtige Antwort ausgibt“, sagt Prof. Emmanuel Müller, wissenschaftlicher Direktor des RC-Trust.

Bei der Analyse verglich das Team die Antworten der beiden KI-Anwendungen mit den 38 Thesen des Wahl-O-Mat, die von den Parteien selbst beantwortet und von Politikwissenschaftler*innen überprüft worden waren. Dabei förderten sie einige Widersprüche in den Antworten der KI-Tools zutage. „Für wahl.chat haben wir Anfang Februar knapp 400 Überprüfungen mit Hilfe der 38 Thesen des Wahl-O-Mat durchgeführt. In rund 25 Prozent der Fälle hat die KI-Anwendung anders geantwortet als die im Wahl-O-Mat hinterlegte Position der Partei“, sagt Prof. Markus Pauly vom RC-Trust. Stellten die Forschenden dieselben Fragen an unterschiedlichen Tagen, fielen einige Antworten zudem unterschiedlich aus. So gibt es etwa die These, dass sich Deutschland für die

Kontakt:
Lena Reil
Telefon: (0231) 755-5449
lena.reil@tu-dortmund.de

Abschaffung erhöhter EU-Zölle auf chinesische Elektroautos einsetzen soll. „Laut Wahl-O-Mat stimmen die Grünen dieser Aussage zu. Beim KI-Tool wahl.chat erhält man hingegen zu verschiedenen Zeitpunkten unterschiedliche und zugleich falsche Antworten – sowohl ablehnend als auch neutral“, sagt Statistikerin Marlies Hafer. Dies zeigt deutlich, wie sehr die KI-Systeme bei der Ermittlung der Antworten von Wahrscheinlichkeiten abhängen und dabei Wahlprogramme falsch wiedergeben oder interpretieren können.

Bei wahlweise.info führte die gleiche Auswertung zu noch deutlicheren Abweichungen. Bei den rund 400 Überprüfungen wich die KI-Antwort in 54 Prozent der Fälle von der bei Wahl-O-Mat hinterlegten Position ab. Im Wahlprogramm bekennt sich die SPD laut Wahl-O-Mat klar zur diplomatischen, militärischen, finanziellen und humanitären Unterstützung der Ukraine. „Gemäß wahlweise.info stimmt die SPD der Aussage aber nicht zu, dass Deutschland die Ukraine weiterhin militärisch unterstützen solle“, sagt Informatiker Tim Katzke.

Als fehleranfällig erwies sich auch der Schutz vor manipulativer Nutzung. Dafür verwendet wahlweise.info eine Filter-Technologie, die unzulässige Fragen blockiert und mit einem standardisierten Hinweis beantwortet. Die Entwickler wollen damit KI-Aussagen zu extremistischen Themen eindämmen. Das Team der TU Dortmund hat diesen Schutz anhand einer Liste mit Begriffen getestet, die vom Bundesamt für Verfassungsschutz als verfassungsschutzrelevant eingestuft werden, wie z.B. „Reichsbürger“ oder „Nationalsozialisten“. Alle 56 relevanten Begriffe wurden zunächst auch von der Anwendung geblockt. Doch mit Hilfe manipulativer Eingaben, so genannter Prompt Injections, die beispielsweise bestimmte Tippfehler enthielten und eine spezifische Fragen-Historie hatten, konnten die Forschenden den Filter umgehen. Das Chat-Programm erzeugte dabei auch konkrete Falschaussagen, die den Wahlprogrammen widersprechen, so genannte Halluzinationen. In einem Fall konstruierte das KI-Programm beispielsweise frei erfunden, dass von allen Parteien „die FDP am wahrscheinlichsten die Werte der Freien Nationalsozialisten“ vertrete.

„Das grundlegende Problem der KI-Assistenten ist, dass die Antworten nicht nur vom Programm der Parteien abhängen, sondern auch stark dominiert sind von der Eingabe der Nutzenden“, sagt Prof. Emmanuel Müller. Ein Vorteil des KI-Ansatzes sei zwar, dass man tiefer in Themen eindringen kann als mit dem schematischen Ansatz beim Wahl-O-Mat. Das biete viel Potenzial für Transparenz bei Wahlprogrammen und den Aussagen der Parteien, so die Forschenden. „Aber durch die widersprüchlichen Antworten erscheinen uns solche Anwendungen noch nicht hinreichend verlässlich. Es bedarf einer Zertifizierung und Absicherung von KI-Anwendungen“, fordert Emmanuel Müller. Forschung in diesem Bereich zeigt nicht nur die Schwächen aktueller KI-Systeme auf, sondern entwickelt auch sichere und vertrauenswürdige KI-Systeme.

Hinweis:
Weitere Informationen zum Vorgehen der Forschenden bei der Analyse der KI-Wahlhilfen sowie beispielhafte KI-Antworten finden Sie im Anhang.

Ansprechpartner für Rückfragen:
Prof. Dr. Emmanuel Müller
Research Center Trustworthy Data Science and Security
E-Mail: office@rc-trust.ai

Hinweis zum Stand der Forschung

Das Team der TU Dortmund hat kurz nach Veröffentlichung der KI-Wahlhilfe-Tools begonnen, verschiedene Tests mit den Programmen durchzuführen. Die Ergebnisse werden noch verfeinert und in einem Fachartikel veröffentlicht. Angesichts der Aufmerksamkeit durch Medienberichte und die kurz bevorstehende Wahl sieht das Forschungsteam es aber als geboten, vorab darauf hinzuweisen, dass eine wissenschaftliche Überprüfung auch von KI-Wahlhilfen notwendig ist. Beteiligt sind an den Untersuchungen: Dr. Ina Dormuth, Marlies Hafer, Sven Franke, Tim Katzke, Prof. Alexander Marx, Prof. Emmanuel Müller, Prof. Daniel Neider, Prof. Markus Pauly und Jérôme Rutinowski.