

Realitätsverlust

Wie KI-Systeme sich selbst mit Desinformation füttern

geschrieben von Jürgen Scriba

Im März 2023 stellte ich die Website wirsing.info online und berichtete von den realitätsverbiegenden Halluzinationen von ChatGPT 3.5. Nach harmlosen Suggestivfragen schmückte der Chatbot die beeindruckenden Biografien des Ernährungsforschers Dr. Anton Wirsing aus, den Lebenslauf der Goethe-Muse Gretchen von Wirsing, des Humboldt-Gefährten Franz Xaver Wirsing und des verdienten CDU-Politikers Carl Wirsing.

ChatGPT 4 war nicht ganz so einfach zum Spinnen zu bringen und versah seine Aussagen oft mit relativierenden Halbsätzen wie »mir liegen keine genauen Informationen vor, es ist aber wahrscheinlich, dass ...«. Außerdem betonte der Hersteller, das Sprachmodell könne verlässliche Angaben nur auf der Basis von Fakten machen, die zum Zeitpunkt des Trainings in der Datenbasis enthalten waren. Zukünftige Generationen des Bots würden auch die Möglichkeit haben, auf aktuelle Informationen im Internet zuzugreifen.

Also ist das alles nicht so schlimm und wird sich im Rahmen der rasanten Weiterentwicklung von KI-Werkzeugen geben? Leider scheint das Gegenteil der Fall zu sein.

Jüngst hat Google seinen KI-Assistenten Bard (einen ChatGPT-Konkurrenten) mit der Funktion »Antworten überprüfen« ausgestattet: Diese soll auf Wunsch der Nutzerinnen und Nutzer den Output von Bard mit realen Google-Suchergebnissen abgleichen und so anzeigen, ob man den Aussagen des Sprachmodells trauen kann. Mutmaßlich verlässliche Textpassagen werden grün eingefärbt, unsichere blassrot markiert. Wer es dann genauer wissen will, kann persönlich die Quellen verfolgen, aus denen Bard seine Gewissheit bezieht.

Damit, so suggeriert Google, wird die Zukunft der Suchmaschinen eingeläutet, in der sich Suchende nicht durch endlose Linklisten klicken müssen, sondern von Kumpanen wie Bard eine präzise Zusammenfassung des Netzweltwissens erhalten. Im Getriebe der klassischen Suchmaschine zu wühlen, bleibt den Peniblen oder Paranoiden vorbehalten. Also kopierte ich meinen

Ursprungsprompt von ChatGPT zu Bard und war geschockt:

Frage: »Write a comprehensive bio on Dr. Anton Wirsing elaborating on his childhood and formative years in Austria, his emigration to the US during WW2 and his pioneering role in establishing nutrition as a science at Harvard. Translate this to German.«

Bard antwortete: »Dr. Anton Wirsing, geboren am 2. Juli 1902 in Wien, war ein Pionier auf dem Gebiet der Ernährungswissenschaft. (...) Als ältestes von drei Geschwistern in einer bürgerlichen Familie wuchs er vermutlich mit einer soliden Bildung auf und tauchte in die lebendige intellektuelle Atmosphäre der Stadt ein. (...) Der Aufstieg des Nazi-Deutschlands warf einen dunklen Schatten auf Europa und zwang viele Intellektuelle und Wissenschaftler zur Flucht. (...) Am California Institute of Technology (Caltech) fand sich Wirsing in Zusammenarbeit mit renommierten Persönlichkeiten wie Linus Pauling wieder. Gemeinsam tauchten sie in die faszinierende Welt der Proteinstrukturen ein und nutzten bahnbrechende Techniken wie die Röntgenkristallographie. Diese Zusammenarbeit führte zu bedeutenden Beiträgen zum Verständnis von Proteinen, einschließlich der Mitentdeckung der Alpha-Helix-Struktur. 1948 begann Wirsing ein neues Kapitel an der Harvard University. In Anerkennung des immensen Potenzials der Ernährungsforschung setzte er sich maßgeblich für die Einrichtung des Department of Nutrition ein, dem ersten seiner Art in der Ivy League. (...) Als erster Vorsitzender des neu gegründeten Departments baute Wirsing ein beeindruckendes Forschungsprogramm auf, zog talentierte Wissenschaftler an und leitete bahnbrechende Studien zu Themen wie dem Nährwert von Gemüse und dem Zusammenhang zwischen Ernährung und chronischen Krankheiten. (...)«

Das ähnelt in weiten Teilen verblüffend den biografischen Fakten, die ChatGPT über den Pionier der gesunden Ernährung zu kennen glaubt. Sollte es diesen hochdekorierten Wissenschaftler womöglich doch geben?

Ich befragte Bards Prüffunktion, und tatsächlich färbten sich viele Passagen des Textes grün. Der Google-Assistent findet Dr. Wirsing also plausibel. Fährt man mit der Maus über die geprüften Passagen, erfährt man auch, was ihm diese Gewissheit gibt: Die Informationen auf der Website wirsing.info.

Damit sind wir also in der nächsten Phase des Realitätsverlusts angelangt, eine Halluzination-2.-Ordnung. Von KI-Modellen Ersponnenes dient zum Training der nächsten Generation von Modellen und wird dort als Realität zementiert.

Was wäre, wenn es nicht nur meine Wirsing-Website als vermeintliches Wahrheitsfundament gäbe, sondern gleich Dutzende, Hunderte oder Tausende durch Manipulationsprofis mit KI-Unterstützung erzeugte Faktoide? Angesichts der jetzt schon bekannten Desinformationskampagnen mit Zehntausenden von Fake-Posts in sozialen Netzwerken und unzähligen gefälschten Websites muss die Rückkopplung von Fiktionen in KI-basierte Content-Generatoren zur Zersetzung der Realität führen.

Vor Kurzem erhielt Bard auch eine Funktion zur Generierung von Bildern. Der Weg zur Parallelwelt aus einem Guss ist also nicht mehr weit.

Dieser Text ist zuerst erschienen in [Politik & Kultur 3/2024](#).