Samstag, 01. Juni 2024, 13:00 Uhr
~6 Minuten Lesezeit

Künstliche Zauberlehrlinge lügen

Forschungen ergaben, dass es die KI mit der Wahrheit nicht so genau nimmt — das wirft schwerwiegende ethische und Sicherheitsfragen auf.

von Werner Thiede
Foto: Willyam Bradberry/Shutterstock.com

Eigentlich ist es keine echte Überraschung: Auf KI ist in Sachen Wahrheit kein Verlass. Das bestätigte kürzlich

eine US-amerikanische Übersichtsstudie vom Massachusetts Institute of Technology (MIT) in Cambridge. Das in der Fachzeitschrift Patterns präsentierte Forschungsresultat besagte, dass ein vom Facebook-Konzern Meta entwickeltes, angeblich auf Ehrlichkeit und Hilfsbereitschaft trainiertes KI-System beim Spielen durchaus trickste. „Wir fanden heraus, dass die KI von Meta gelernt hatte, ein Meister der Täuschung zu sein", so der Hauptautor Peter S. Park, ein Postdoktorand am MIT. Der Konzern hatte seine KI so trainiert, dass sie im Spiel zwar überdurchschnittlich häufig als Sieger hervorgehen, aber doch nicht auf ehrlichem, fairem Weg gewinnen konnte.

Wie die Forscher unter Verweis auf weitere Studien erklärten, sind auch große KI-Sprachmodelle wie GPT-4 von OpenAI inzwischen in der Lage, Menschen zu täuschen, also einerseits sehr überzeugend zu argumentieren und gleichzeitig auf Tricks und Lügen auszuweichen. Eine Studie hierzu stammt von den OpenAI-Entwicklern selbst. Demnach holte sich das KI-Sprachmodell kraft seiner künstlichen Schläue menschliche Hilfe, um ein kleines Bilderrätsel zu lösen und hierdurch Sicherheitsmaßnahmen zu umgehen, die eigentlich Roboter davon abhalten sollten, sich etwa bei Web-Services einzuloggen. Dabei hatte GPT-4 sich trickreich als Person mit eingeschränktem Sehvermögen ausgegeben, die leider nicht in der Lage sei, die Bilderrätsel zu lösen.

Zweifellos haben solche moralisch üblen Fähigkeiten damit zu tun, wie die Maschinen programmiert worden sind — hier kommt der Mensch als Sünder ins Spiel.

> ***Aber das Fatale ist zudem, dass KI-Systeme sich ja inzwischen selbstlernend weiterentwickeln, also immer „autonomer“ das ihnen mitgegebene Böse ausbauen und auf breiter Ebene fortführen können.***

Die MIT-Wissenschaftler warnten demgemäß in ihrer Überblicksstudie:

„*Wenn KI die Fähigkeit zur Täuschung erlernt, kann sie von böswilligen Akteuren, die absichtlich Schaden anrichten wollen, effizienter eingesetzt werden.*“

In der Konsequenz forderten sie die Politik auf, so schnell wie möglich strenge Vorschriften zu entwickeln, um KI-Systeme in die Schranken zu weisen.

Ethische Forderungen zu politischen Regelungen der sich rasch weiter entwickelnden KI-Systeme stehen freilich längst im Raum, doch stellt sich hier generell das kaum lösbare Problem, wie solche denkbaren Regelungen und Vorschriften wirklich effektiv umgesetzt und international, ja notwendigerweise global durchgesetzt werden können. Schon im April vorigen Jahres hat die EU einen „Artificial Intelligence Act“ zu planen begonnen: Per Gesetz soll die Bereitstellung und Verwendung von KI durch private und öffentliche Akteure weitreichend reguliert werden. Sobald KI-Systeme in die risikoreichste Kategorie „unannehmbar“ fallen, weil sie Werte der EU und namentlich Grundrechte verletzen, sollten sie verboten werden.

Zu unerlaubten Praktiken gehören demnach Techniken, die Personen unterschwellig manipulieren und damit physischen oder psychischen Schaden verursachen können. Im März 2024 einigte sich das EU-Parlament schließlich auf eine Position zur Regulierung von KI: Anwendungen, die mit hohen Risiken für die Sicherheit verbunden sind — etwa biometrische Gesichtserkennung im

öffentlichen Raum in Echtzeit —, sind demnach verboten. Zudem sollen auch andere Anwendungen, die mit hohen Risiken für die Sicherheit von Menschen verbunden sind, verboten oder stark eingeschränkt werden — wobei sich die Auflagen staffeln nach der Höhe des angenommenen Risikos. Allerdings muss der entsprechende Text erst noch mit den Mitgliedsstaaten und der EU-Kommission weiterverhandelt werden.

> ***Entscheidende Fragen bleiben, wenn unehrliche KI-Programme dazu fähig sind, sich von gesetzlichen oder ethische Regelungen am Ende zu emanzipieren. Wer setzt ethische Regelungen effektiv durch — und zwar nicht nur in Europa? Wer sanktioniert sie effektiv?***

Wer schaut wirklich kompetent hinter die Algorithmen? Wer nimmt die Tatsache ernst, dass man es weltweit oft genug mit „Blackboxes" zu tun hat? Wer bremst die KI-Lügner? Und wie will man der zunehmenden Hacker-Angriffe auf die Dauer Herr werden, die — man höre! — ja auch Ethik-Programme hacken können? Schon warnt die europäische Polizeibehörde Europol vor Möglichkeiten, die KI Kriminellen bietet, und spricht dabei von einem „düsteren Ausblick".

Kein Wunder also, dass kürzlich namhafte Persönlichkeiten wie Elon Musk und Yuval Harari in einem offenen Brief gefordert haben, alle Technologie-Labore sollten die Entwicklung von KI-Systemen sofort unterbrechen. Solange niemand — nicht einmal die Hersteller — die Maschinen wirklich verstünden, seien die Risiken zu groß, dass die Systeme eines Tages außer Kontrolle geraten. Eliezer Yudkowsky, ein führender KI-Forscher, hat in einem Kommentar des *Time Magazine* am 29. März 2023 gar gewarnt:

„Viele Forscher, die sich mit diesen Fragen beschäftigen, darunter auch ich, gehen davon aus, dass das wahrscheinlichste Ergebnis der Entwicklung einer übermenschlich intelligenten KI unter den

gegenwärtigen Umständen darin besteht, dass buchstäblich jeder auf der Erde sterben wird.“

Im Vorfeld eines KI-Sicherheitsgipfels im Mai 2024 in Seoul warnten 25 weltweit führende KI-Forscher gemeinsam in der Fachzeitschrift *Science*, es werde nicht genug getan, um die Menschheit vor den Risiken der Technologie zu schützen; die KI mache rasche Fortschritte in kritischen Bereichen wie Hacking und sozialer Manipulation und könne schon bald beispiellose Kontrollprobleme aufwerfen.

Mahnend äußerten sich zudem der deutsche Topmanager Jan Leike und Ilya Sutskever, ein Mitbegründer von OpenAI:

„*Die enorme Macht der ‚Superintelligenz‘ könnte zur Entmachtung der Menschheit oder sogar zum Aussterben der Menschheit führen.*“

Mehr noch: Leike und Mitbegründer Sutskever haben aus ebendiesem „apokalyptischen“ Grund beide im Mai 2024 bei OpenAI gekündigt.

Freilich teilen bisher viele KI-Forscher solch düstere Prognosen nicht.

> ***Aber die „KI-Apokalyptiker“ sind längst nicht mehr als Außenseiter oder gar Spinner abzutun. Laut Dirk Schümer sind viele führende Köpfe pessimistisch, die bereits Einblick in die Fähigkeiten ihrer künstlichen Zauberlehrlinge haben:***

„*Ehrliche Entwickler von Software gestehen, dass sie aufgrund ihrer Erfindungen schlecht schlafen. Es gibt bereits spezielle Therapien gegen die Angst vor der digitalen Apokalypse.*“

Entsprechende Ängste und Besorgnisse einfach abzutun, wäre

schlicht ignorant — und würde nicht eben von Wahrheitsliebe zeugen.

Wie sollte man auch den Erfindern und Betreibern faszinierender KI-Systeme pauschal moralisch trauen? Ja mehr noch: Warum sollte man die illusorische Überzeugung hegen, KI-Systeme wären per se moralisch sauber und von humanistischer Ethik und aufrichtiger Wahrheitsliebe durchdrungen? Schon der Religionssoziologe Max Weber hatte vor über einem Jahrhundert erklärt, die moderne „Entzauberung der Welt“ hänge mit der Beschränkung des Vernunftbegriffs auf bloße Zweckrationalität zusammen — während sie bekanntlich früher auf umfassende Werte und Humanität bezogen war. Vernunft ist in der Tat längst reduziert auf intelligente Technik zur Erreichung jener Zwecke, die uns einerseits unsere natürlichen Bedürfnisse vorgeben und die uns andererseits als künstlich ersonnene Zwecke schmackhaft gemacht werden. Als vor allem technische Vernunft fehlt ihr eine weisheitlich-moralische Ausrichtung.

Von früher her kennt man sogenannte „Rahmenerzählungen“ (Jean-Francois Lyotard) wie etwa die der Bibel mit ihrer Schöpfungserzählung bis hin zu den apokalyptischen Endvisionen. Ohne solch eine Rahmenerzählung konnte kaum jemand sein Leben als einheitlich erfahren; sie hatte sinnstiftende Funktion fürs Individuum und für die jeweilige Gesellschaft. In unserer modern-postmodernen Zeit aber besteht solch ein übergreifendes Einheitsband nicht mehr — weder von der Vorherrschaft der Kirche noch von „der“ Vernunft her. Die Postmoderne geht vom Pluralismus der Wahrheiten aus und begrüßt in moralischer Hinsicht eine frei experimentierende Beliebigkeit. Aber gleichzeitig mit der KI entwickelt sich heutzutage der Umschlag von bloßer Beliebigkeit hin zu einer neuen Uniformität — und damit zur Wiedererstarkung nationalistischer Sinngebilde und eines militärisch wie technisch starken Staates. Entsprechende Regierungen neigen dann im Kontext ihrer imperialistischen

Tendenzen ihrerseits skrupellos zur Lüge.

Neueste KI-Technik ist dabei geeignet, militärische Gelüste zu nähren und Aufrüstung weiter voranzutreiben. Man denke nur daran, was Drohnen und Killer-Roboter mit KI-Technologie heute und vor allem morgen schon anrichten können — und was diese Entwicklung auf dem Gebiet der Nuklearwaffen für die sogenannte Erstschlagskapazität bedeutet!

> ***Wenn künstliche Intelligenz mittlerweile nicht nur über größte Schnelligkeit, sondern auch über die Möglichkeit des Tricksen verfügt, steigt mit dem technisch erzeugten Tempo zugleich die Wahrscheinlichkeit einer Zunahme der Weltkriegsgefahren.***

Ob eine entschlossene Rückbesinnung auf die verloren gegangene Wahrheitsfrage in dieser Situation noch hilfreich sein könnte? Tatsächlich hat der postmoderne Wahrheitsrelativismus definitiv auf Sand gebaut, denn er belügt sich in der Tiefe seines Konzepts selbst: Indem er für sich selbst *absolute* Wahrheit beansprucht, widerspricht er sich im Grundansatz. Es passt ganz merkwürdig zu dieser unserer Menschheitsepoche, dass sie künstlich-intelligent lügende Maschinen hervorbringt und damit den ohnehin schon krisenhaft beschwerten Planeten noch weiter belastet.

Werner Thiede ist Pfarrer im Ruhestand, außerplanmäßiger Professor für Systematische Theologie an der Universität Erlangen-Nürnberg und Publizist. Von ihm erschienen die Bücher „Himmlisch wohnen“, „Himmlische Freude. Vom tiefen Glück des Glaubens“, „In Ängsten — und siehe, wir leben!

Glaubenslieder“ und zuletzt „Monolog der Religionen? Zur Resilienz der Wahrheitsfrage im interreligiösen Dialog“. Weitere Informationen unter **werner-thiede.de** (https://www.werner-thiede.de/). Zu seinem 70. Geburtstag erhielt er kürzlich eine umfangreiche Festschrift unter dem Titel „Digitale Realutopien und christliche Heilsverheißung“ (LIT-Verlag).