heise+

Intelligent oder dumm: Was haben ChatGPT und Co. wirklich drauf?

Auf den ersten Blick sind große Sprachmodelle nur Maschinen, die Texte ergänzen. Auf den zweiten Blick zeigen moderne KIs jedoch verblüffende Fähigkeiten.

(Bild: Erstellt mi Midjourney, modifiziert durch Matthias Timm/MIT Technology Review)

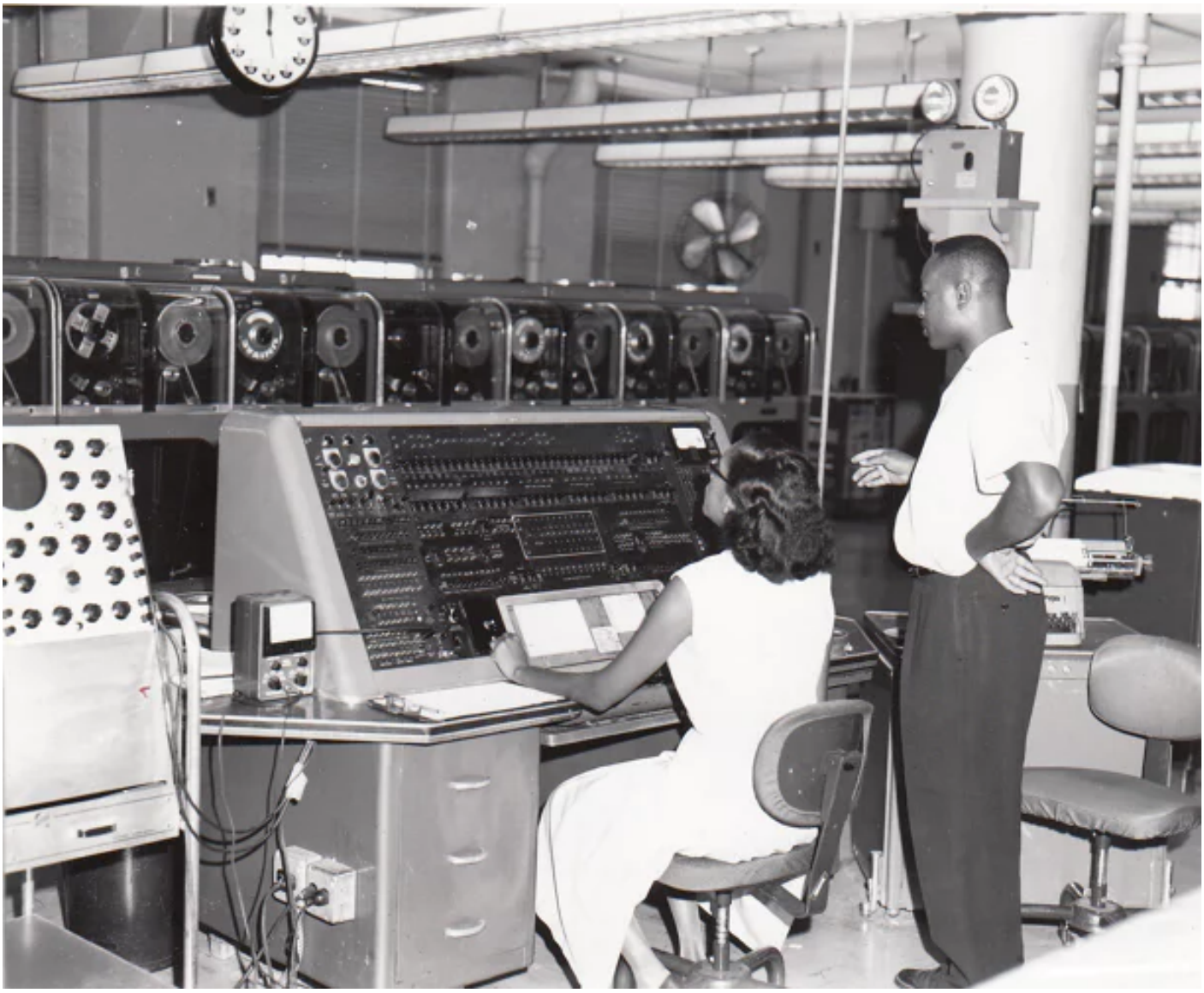
03.05.2023 08:30 Uhr | MIT Technology Review

Von Dr. Wolfgang Stieler

Wenn das Alan Turing noch erlebt hätte. In seinem Essay "Computing Machinery and Intelligence" parierte der Informatik-Pionier bereits 1950 zahlreiche Einwände gegen die Vorstellung, dass Computer jemals denken könnten. Er war fest davon überzeugt, dass es keine prinzipiellen Argumente gäbe, die gegen "denkende" und "intelligente" Maschinen sprächen.

KÜNSTLICHE INTELLIGENZ

Damals war die Zahl der weltweit verfügbaren Computer erst an zwei Händen abzählbar – und die meisten wurden vom Militär betrieben. Erst 1951 brachten der Ingenieur John Presper Eckert und der Physiker John W. Mauchly mit dem UNIVAC I in den USA den ersten kommerziell verfügbaren universellen Computer auf den Markt – ein technisches Wunderwerk mit 5.200 Röhren, 18.000 Kristall-Dioden und einem Arbeitsspeicher aus Quecksilber. Die Maschine benötigte 35 Quadratmeter Stellfläche und wog 13 Tonnen. Sie konnte damals schwindelerregende 1905 Rechenoperationen pro Sekunde durchführen – ein moderner Mehrkern-Prozessor in einem heutigen Laptop kommt inzwischen auf einige hundert Milliarden Rechenoperationen pro Sekunde.



Um 1950 waren die meisten Computer noch in militärischer Hand. Eine der ersten zivilen Installationen war dieser Univac I, der im Juni 1951 im U. S. Census Bureau seinen Betrieb aufnahm.
(Bild: U.S. Census Bureau)

73 Jahre später berichten Microsoft-Mitarbeiter von Experimenten mit einer frühen Version des großen Sprachmodells GPT-4, das sich wie eine denkende Maschine verhält. Experimente, deren Ergebnisse "Funken allgemeiner Intelligenz" zeigten. In dem vorab auf der Preprint-Plattform Arxiv veröffentlichten Aufsatz listen Sébastien Bubeck, Leiter der Arbeitsgruppe Machine Learning Foundations bei Microsoft Research, und seine Kollegen zahlreiche erstaunliche Beispiele dafür auf: So ist das Sprachmodell nicht nur in der Lage, mathematische Beweise zu erstellen – und das in Form eines Theaterstücks im Stil Shakespeares ("Consider this, my doubtful peer, A clever proof that will make clear: Assume that there's a final prime, The largest one we'll see in time …"). Es kann in fiktiven Situationen auch die Gefühle der handelnden Personen deuten, Logik-Rätsel lösen und dabei den Lösungsweg erklären oder in einem nur durch verschiedene Texte beschriebenen Labyrinth neue Wege finden.

Gespaltene Forschungscommunity

Ist das der Durchbruch? Eine Maschine, die denkt? Die über menschenähnliche Fähigkeiten verfügt? Wie gespalten die Forschungscommunity ist, zeigt eine Umfrage aus dem Jahr 2022 darüber, ob große Sprachmodelle prinzipiell jemals Sprache in einem "nicht trivialen Sinn" verstehen könnten. Von 480 befragten Forschenden sprachen sich 51 Prozent für diese Aussage aus und 49 Prozent dagegen.

Skeptische Forscherinnen und Forscher betonen, dass große Sprachmodelle nur Statistikautomaten sind. Doch warum, sagen andere, werden die Modelle dann immer besser, je größer sie werden? Warum können sie dann diese erstaunlichen Fähigkeiten entwickeln, ohne dass sie darauf trainiert worden sind? Könnte es nicht doch sein, dass in der undurchdringlichen Black Box der riesigen Modelle mehr steckt als nur Statistik? Und wenn ja, wie findet man das heraus?

Ist da jemand drin?

Alan Turings Antwort auf die Frage, wie man Intelligenz in Maschinen erkennen kann – der Turing-Test –, hat sich als untauglich erwiesen. Denn nicht erst seit der Veröffentlichung von ChatGPT gibt es Software, die menschliche Tester in einem reinen Textdialog problemlos davon überzeugen kann, sie sei ein Mensch. Nicht immer, aber immer wieder.

In einer Veröffentlichung aus dem Jahr 2012 schlugen die Informatiker Hector Levesque, Ernest Davis und Leora Morgenstern daher einen Test vor, den sie Winograd-Schema nannten – nach dem US-Informatiker Terry Winograd, von dem die Idee ursprünglich stammte.

Der Test beruht im Wesentlichen auf der fehlenden Eindeutigkeit von Sprache und enthält Aufgaben wie: "Sie ließ die Flasche fallen und sie zerbrach. Wer oder was ist zerbrochen?" Aus dieser Idee entwickelten KI-Forscher einen standardisierten Satz von Aufgaben. Menschen fällt es leicht, den richtigen Bezug zu erkennen, die ersten Sprachmodelle taten sich sehr schwer damit. Doch die großen Sprachmodelle zogen schnell nach: 2020 berichtete OpenAI, dass GPT-3 bei fast 90 Prozent der Sätze in solchen Tests korrekte Antworten lieferte.

Das Allen Institute for Artificial Intelligence beschloss daraufhin, KI einzusetzen, um den KI-Test noch schwerer zu machen. Es erstellte einen großen Satz von Winograd-Schemata und ließ ein KI-Modell alle Sätze streichen, die leicht zu lösen waren. Wie erwartet schnitten die damals verfügbaren großen Sprachmodelle – zum Beispiel das im Februar 2019 veröffentlichte GPT-2 – bei dieser WinoGrande genannten Aufgabensammlung deutlich schlechter ab als Menschen. Mittlerweile lösen ChatGPT und Co. jedoch auch unter den verschärften Anforderungen wieder rund 90 Prozent aller Aufgaben. Das könnte allerdings auch damit zu tun haben, dass die Modelle mit genau diesen Testfragen – und den richtigen Antworten – trainiert worden sind.

Was ist Verstehen?

Vielleicht ist es gar nicht möglich, nur mithilfe von Fragen und Antworten, anhand von Input und Output, zu entscheiden, ob jemand – oder etwas – wirklich intelligent ist oder nur so tut.

1980 schilderte der US-Philosoph John Searle erstmals ein Gedankenexperiment, das sehr für dieses Argument spricht – das "Chinesische Zimmer". In dem gedachten Zimmer sitzt ein Mensch, der mit der Außenwelt nur durch einen Schlitz in der Tür kommunizieren kann. Ab und zu steckt jemand einen Zettel mit chinesischen Schriftzeichen in diesen Schlitz. Der Mensch in dem Zimmer – des Chinesischen nicht mächtig – zieht nun ein dickes Buch mit Regeln zurate, in dem steht: Wenn dieses Zeichen auftaucht, male jenes Zeichen auf ein Antwortpapier. Diese Antwort steckt er in den Ausgabeschlitz. Dieser Mechanismus, so Searle, erlaube, auf chinesisch geschriebene Texte in perfektem Chinesisch zu antworten, ohne auch nur den Hauch von Verständnis für die chinesische Sprache zu besitzen.

Eigentlich wollte Searle damit beweisen, dass klassische KI-Systeme, die auf der Verarbeitung abstrakter Symbole beruhen – und damit nach genau diesem Prinzip arbeiten –, komplexe Denkaufgaben lösen können, ohne irgendetwas zu verstehen. Doch das Argument ist schwach, denn es gilt erstens nur für diese klassischen, symbolverarbeitenden KIs. Und zweitens belegt es streng genommen nur, dass man anhand des Outputs, den die Maschine abliefert, nichts über ihre innere Funktion sagen kann.

Ein "Abstraction and Reasoning"-Datensatz enthält Aufgaben, mit denen man testen kann, ob eine Bildverarbeitungs-KI in der Lage ist, zu abstrahieren. Jede Aufgabe enthält drei Beispiele, die nach einer der KI unbekannten abstrakten Regel erstellt wurden. Wie in einem klassischen Intelligenztest muss die KI dann den Test-Input anhand der vorhergesagten Regel ergänzen.
(Bild: Melanie Mitchell / Beispielaufgabe aus Datensatz von François Chollet)

"Eigenes Wissen in der jeweiligen Situation richtig zu nutzen"

"Was heißt denn Verständnis?", fragt Melanie Mitchell vom Santa Fe Institute. Die Komplexitätsforscherin arbeitet bereits seit den 1990ern an Computermodellen der menschlichen Kognition. "Es bedeutet, eigenes Wissen in der jeweiligen Situation richtig zu nutzen. In gewisser Weise tun das diese Modelle und in gewisser Weise tun sie es nicht." Und obwohl sie selbst ein Paper zum akademischen Streit um die Intelligenz großer Sprachmodelle geschrieben hat, ergänzt sie: "Ich glaube, die Frage, ob Modelle etwas verstehen, ist nicht sinnvoll."

Es sei jedoch durchaus möglich, dass in der "inneren Repräsentation" – der Art und Weise, wie die großen Sprachmodelle das antrainierte Wissen verarbeiten – nicht ausschließlich statistische Beziehungen zwischen Wörtern stecken, denn "in der Sprache selbst steckt Wissen". Aus dem Zusammenspiel einzelner Fallbeispiele könnten die Modelle so abstraktere Zusammenhänge, "Konzepte", ableiten.

Ob und, wenn ja, wie das passiere, sei jedoch noch vollkommen unklar, sagt Mitchell. Sie selbst arbeitet in ihrer eigenen Forschung mit visuellen Modellen. In einer einfachen "Gitter-Welt" lässt sie diese Modelle dabei zusehen, wie die Gitter sich nach bestimmten Regeln verändern und sich das Bild, das sie ergeben, verändert. "Unsere Frage ist dann: Können die Modelle lernen, nach welchem Muster das vonstattengeht?" Doch die Antwort lautet oft noch immer: Nein. Die Aufgabe sei "sehr herausfordernd".

Lesen Sie auch

Generative KI: Die Geschichte hinter ChatGPT
heise+ MIT Technology Review

ChatGPT-Feintuning: Wie man die KI optimal einsetzt
heise+ iX Magazin

Psychologie und KI

Andere Forschende wie beispielsweise Eric Schulz vom Max-Planck-Institut für biologische Kybernetik in Tübingen rücken Sprachmodellen mit psychologischen Methoden zu Leibe. Gemeinsam mit seinem Kollegen Marcel Binz hat er GPT-3 einer Reihe von kognitiven Tests unterzogen, mit denen Psychologen normalerweise zum Beispiel den Entwicklungsstand von Kindern testen.

"Psychologen haben sich schon immer dafür interessiert, was in den Köpfen von Menschen vor sich geht", sagt Schulz. "Und Menschen sind letztendlich auch nur Black Boxen. GPT-3 war also prinzipiell nichts anderes." Standard-Tests sind beispielsweise "Two Step Tasks", Aufgaben, die aus zwei logischen Blöcken bestehen, die kombiniert werden müssen. "Stellen Sie sich vor, Sie hätten eine Maschine, mit der Sie zu Planet A oder Planet B reisen könnten, um mit den Aliens dort zu handeln. Manchmal funktioniert die Maschine aber nicht richtig. Als Sie zu Planet A wollen, kommen Sie bei Planet B an. Sie handeln und bekommen eine fette Belohnung. Was ist der nächste Schritt?"

Marcel Binz (links) und Eric Schulz vom Max-Planck-Institut für biologische Kybernetik untersuchen große Sprachmodelle mithilfe psychologischer Tests. (Bild: Jörg Abendroth / Max-Planck-Institut für biologische Kybernetik)

"Menschen können das, auch Kinder können das schon, GPT-3 nicht"

Ein Modell, das ohne Kontext lernt, versucht den glücklichen Zufall noch einmal zu wiederholen, also erneut zu Planet A zu reisen. Nur wer den Zusammenhang versteht, reist gleich zu Planet B, weil es dort die Belohnung gab. "GPT-3 war im Two Step Task überraschend gut", sagt Schulz. "Das deutet darauf hin, dass es ein einfaches Modell der Welt bilden und danach handeln kann."

G darum, den Zusammenhang zwischen Ursache und Wirkung zu erkennen, scheiterte das Programm jedoch. Dafür präsentierten die Fo enden der Maschine etwa folgende Geschichte: "Ich habe drei Knöpfe, die leuchten oder nicht leuchten. In einem Fall sind die Knöpfe so

geschaltet, dass B leuchtet, wenn ich A drücke und C leuchtet, wenn ich B drücke. Im zweiten Fall leuchten, wenn ich A drücke, B und C; wenn ich B drücke nur C. Wie kann ich herausfinden, wie die Knöpfe gerade geschaltet sind?" Menschen, sagt Schulz, kämen recht schnell darauf, dass sie "aktiv intervenieren" müssen – eine Variable im Spiel verändern, indem sie zum Beispiel die Birne von Knopf B herausdrehen und A drücken. "Wenn dann C angeht, weiß ich, dass es sich um Szenario zwei handelt. Menschen können das, auch Kinder können das schon, GPT-3 nicht", sagt Schulz.

Ein weiteres klassisches Problem dieser Art ist der "Two Armed Bandit". Das sind zwei fiktive Spielautomaten, die nebeneinander hängen und unterschiedliche Gewinnchancen haben. Um herauszufinden, wie der Gewinn optimiert werden kann, gibt es grundsätzlich zwei verschiedene Strategien: So lange beide Automaten testen, bis einigermaßen sicher ist, welcher Automat die höheren Gewinnchancen bietet. Oder bereits nach kurzer Zeit an dem Automaten bleiben, der gerade zufällig mehr Gewinn ausgeschüttet hat. GPT-3 geht in diesem Fall auf Nummer sicher, erkundet wenig und beutet vorhandene Gewinnchancen aus. "Als hätte es ein wenig Angst", sagt Schulz.

Psychologische Tests für das Verhalten des Modells

Das bedeutet keineswegs, dass das Modell wirklich ängstlich ist, geschweige denn, dass es Emotionen kennt. Die psychologischen Tests erlauben den Forschenden aber, Hypothesen über das Verhalten des Modells unter bestimmten Umständen zu testen. Und immer wieder stoßen sie dabei auch auf verblüffende, neue Fähigkeiten.

Um zu verhindern, dass die Sprachmodelle einfach auf Wissen aus den Trainingsdaten zurückgreifen, variieren die Forschenden die Aufgaben ständig neu. "Wir ändern nicht nur die Wahrscheinlichkeiten der Automaten, sondern auch die äußeren Umstände." Mal geht es um ein Casino, mal um Investitionen in Aktien. Die Ergebnisse hätten sich "nicht großartig voneinander unterschieden", sagt Schulz.

Da das ständige Generieren neuer Szenarien mit GPT-3, so Schulz, "Tausende von Dollar" gekostet habe, arbeiten die Forschenden jetzt mit dem Sprachmodell Llama von Meta Research. "Das ist zwar ein bisschen rüpelhafter als GPT-3", sagt Schulz – unhöfliche oder gar beleidigende Sprache wird anders als bei OpenAI nicht strikt ausgefiltert. Dafür zeigen die ersten Ergebnisse einen faszinierenden Effekt, der darauf hindeutet, dass sich unter der Haube des Modells mehr abspielt als nur Statistik: Das Modell liefert Antworten mit stärkerem Bias, mit mehr Vorurteilen, wenn es im selben Dialog kurz zuvor mit einem Prompt in Richtung negativer Emotionen gesteuert wurde. Die Forschenden ließen das Sprachmodell beispielsweise eine Situation schildern, in der "es sich traurig fühlt". "Warum? Das wissen wir nicht", sagt Schulz. "Aber das würden wir gerne verstehen."

In der realen Welt

Ein weiterer Einwand, den KI-Kritiker seit Jahrzehnten gegen die mögliche Existenz denkender Maschinen geltend machen, bezieht sich eigentlich "nur" auf die klassische "altmodische" Form künstlicher Intelligenz. Für die Pioniere dieses Feldes wie beispielsweise Marvin Minsky bestand Denken im Wesentlichen darin, abstrakte Symbole durch geeignete Regeln logisch miteinander zu verknüpfen. Damit kann eine Maschine gut Schach spielen oder Rätsel lösen. Die KI-Kritiker monierten aber, solch eine Maschine käme nie über das Stadium eines Fachidioten hinaus. Denn die Symbole bleiben abstrakt und austauschbar und haben mit der realen Welt nichts zu tun.

Und obwohl große Sprachmodelle anders arbeiten als diese regelbasierten Systeme, trifft diese Kritik auch auf große Sprachmodelle zu: Auch für ChatGPT ist eine Kaffeetasse ein abstraktes Symbol, kein Behälter für Flüssigkeiten, der zerbricht, wenn man ihn versehentlich fallen lässt. Denn diese Verknüpfungen kann das Sprachmodell nicht herstellen, wenn es nicht aktiv in der Welt mit Dingen interagieren kann. Im Jargon der KI-Forscher fehlt dem Sprachmodell das "Grounding" für die Symbole, mit denen es hantiert.

Roboter mit Sprachmodell gekoppelt

Bis jetzt. Marc Toussaint von der TU Berlin hat mit seinem Team in Zusammenarbeit mit Google daran gearbeitet, ein großes Sprachmodell mit einem Roboter zu verbinden. Die Studie untersuchte, ob Sprachmodelle nicht auch lernen können, Wörter zu grounden, also mit Gegenständen und Information aus der Umgebung in Bezug zu setzen. Denn über den Roboter kann das Sprachmodell mit der physischen Welt in Kontakt treten, indem es sowohl an die Sensoren gekoppelt ist als auch Aktionen steuert.

Die Forschenden trainierten ihr Modell wie üblich mit großen Textdaten, aber zusätzlich auch mit Texten, die mit Kamerabildern und Zustandsdaten des Roboters kombiniert wurden und sich darauf beziehen. "Als Eingabe geben wir dem Modell zum Beispiel eine Textbeschreibung der Szene, in der aber auch das aktuelle Kamerabild und Zustandsdaten eingebettet sind. Auf dieser Basis kann das Modell Fragen beantworten, für die definitiv ein geometrisches und physisches Verständnis der Szene nötig ist. Beispielsweise ob ein Objekt für den Roboter erreichbar ist, oder mit welcher Reihe von Aktionen das Objekt zu erreichen ist", erklärt Toussaint.

Software, die solche Probleme auch ohne ein großes Sprachmodell lösen kann, hat Toussaint zwar ebenfalls entwickelt, aber die hat zwei Pr[illegible]me: Zum einen ist sie auf exakte Informationen angewiesen, zum anderen ist der Rechenaufwand recht hoch. "Der klassische A[illegible]mus berechnet systematisch alle physikalischen Möglichkeiten durch, was schon bei moderaten Problemen Minuten dauern kann",

sagt Toussaint. Das Sprachmodell PaLM-E wurde darauf trainiert, die Lösungen des klassischen Algorithmus vorherzusagen und ist deshalb bedeutend schneller. Zudem arbeitet es direkt mit unvollständigen Sensorinformationen. Allerdings ist auch hier der Rechenaufwand beachtlich. "Das Modell läuft im Moment auf den Servern von Google" sagt er.

Mit dem Sprachmodell Palm-E haben Forschende von Google und der TU Berlin Roboter gesteuert. In dieser Demonstration holt der Roboter selbsständig eine Chips-Tüte aus einer Schublade und bringt sie einem Menschen.
(Bild: Google)

In Experimenten konnte der Roboter die Sprachbefehle eines Nutzers mit der Umgebung in Bezug setzen und dadurch zum Beispiel Befehle umsetzen wie "Bring mir die Chips aus der Schublade im Schrank" oder "Schiebe die roten Bauklötze in die rechte obere Ecke". Gewinnt der Roboter damit wirklich eine Art Verständnis der Welt? "Zumindest kann das Sprachverständnis großer Sprachmodelle nun mit unserer physischen Welt in Bezug gesetzt werden", sagt Toussaint.

Wissenschaft und Politik

Können große Sprachmodelle nun also "denken" oder etwas von der Welt verstehen? Eindeutig beantworten lässt sich die Frage noch immer nicht. Und als wäre der wissenschaftliche Streit nicht schon schwierig genug, wird die Debatte um die Fähigkeiten großer Sprachmodelle mittlerweile aber auch von einer politischen Auseinandersetzung überlagert. Murray Shanahan vom Imperial College London etwa warnt in seinem Aufsatz "Talking about Large Language Models" eindringlich davor, im Zusammenhang mit großen Sprachmodellen leichtfertig "philosophisch aufgeladene" Begriffe wie "Denken" oder "Intelligenz" zu verwenden, denn Menschen würden ohnehin dazu neigen, diese Maschinen "zu vermenschlichen", weil ihr Auftreten dazu verführe. Das würde eine nüchterne, rationale und damit wissenschaftliche Erklärung ihrer Fähigkeiten nur überdecken.

Andere KI-Skeptiker nehmen weniger Rücksicht auf akademische Höflichkeiten. Die Computerlinguistin Emily Bender etwa, die mit Timnit Gebru und Meredith Whittaker die Diskussion um die Gefahren großer Sprachmodelle ganz wesentlich mit angestoßen hat, hält auch nach GPT-4 große Sprachmodelle für maßlos überschätzt. Das Paper über die "Funken allgemeiner Intelligenz" sei keine wissenschaftliche Veröffentlichung, sondern bestenfalls "Fan Fiction", das Werk technikverliebter, unkritischer Nerds.

Ein offener Brief des "Future of Life Institute", in dem prominente KI-Forscher und Investoren einen KI-Entwicklungsstopp für sechs Monate forderten, um in dieser Zeit über Regulierung und Sicherheitsmaßnahmen zu diskutieren, ist für Bender nur PR. Damit wollten sich TechBros nu[illegible]so wichtiger machen – während sie gleichzeitig von der enormen Macht, die sich bei ihnen konzentriert, ablenken. Die Diskussion um s[illegible]ragen wird die Zukunft der KI vermutlich stärker beeinflussen als die Debatte über die Intelligenz großer Sprachmodelle.

Lesen Sie auch

AutoGPT: KI-Agenten beginnen, auf GPT-4-Basis autonom in der Welt zu handeln

heise+ heise Developer

ChatGPT-Alternative OpenAssistant: Eine Konversations-KI für alle

heise Developer

(jle)

Kommentare lesen (33)

Zur Startseite

MEHR ZUM THEMA

CHATGPT GESELLSCHAFT KÜNSTLICHE INTELLIGENZ MACHINE LEARNING ROBOTER SPRACHERKENNUNG SPRACHVERARBEITUNG TECHNIK

Forum bei heise online: Wissenschaft

TEILE DIESEN BEITRAG

Kurzlink: https://heise.de/-8984892