Von künstlicher Intelligenz erzeugte Bilder sind kaum von echten Fotos zu unterscheiden. Dieses Bild hat die Software Midjourney erzeugt, und zwar anhand der Steuerbefehle „Sigmund Freud behandelt einen modernen Cyborg auf einer Liege“, „in einer Psychotherapie-Sitzung“ und „hyperrealistisch“.

BILD: KI BILD MIDJOURNEY | ERSTELLT VON GESINE BORN | BILDERINSTITUT

KÜNSTLICHE INTELLIGENZ AUF DER COUCH

TEXT: UTE EBERLE

Seit der Veröffentlichung von ChatGPT Ende 2022 wird intensiv darüber diskutiert, ob die künstliche Intelligenz bereits menschenähnliche Denkfähigkeiten besitzt. Eric Schulz vom Max-Planck-Institut für biologische Kybernetik in Tübingen untersucht mithilfe psychologischer Tests, ob dieser Algorithmus Anzeichen einer allgemeinen Intelligenz aufweist.

Eric Schulz ist Kognitionswissenschaftler, er ist also an Denkprozessen im Gehirn des Menschen interessiert. Derzeit beschäftigt er sich mit dem Innenleben von Intelligenz, die ebendieses Gehirn erschaffen hat. „Ich wollte immer schon verstehen, wie Menschen ticken. Jetzt frage ich mich: Wie tickt eigentlich künstliche Intelligenz?“ Um das herauszufinden, unterzieht Schulz die künstliche Intelligenz (KI) klassischen kognitionswissenschaftlichen Experimenten. Seine Ergebnisse haben in ihm die Erkenntnis reifen lassen, dass wir Vorsichtsmaßregeln einführen sollten, wenn wir solche Systeme in unseren Alltag integrieren. Zum Beispiel, dass immer erkennbar sein muss, wann sie zum Einsatz kommen. Nicht weil er glaubt, dass künstliche Intelligenz die Macht übernimmt – eine Sorge, die er oft spürt, wenn er sich mit anderen Menschen über seine Forschung unterhält. Schulz hat da wenig Bedenken. Kritisch sieht er eher die Geheimniskrämerei der Unternehmen, die künstliche Intelligenz entwickeln.

Schulz’ Erfahrungen basieren bisher vor allem auf GPT-3, also dem System, das bis Mitte vergangenen Jahres eines der fortschrittlichsten war. GPT-3 arbeitet noch ohne die Chat-Komponente von ChatGPT und ohne Bilder, die GPT-4 nutzt. Diese beiden Programme wurden in den vergangenen Monaten kurz hintereinander öffentlich gemacht, begleitet von Konkurrenzversionen wie Bard von Google. Doch all diese Systeme folgen demselben Grundprinzip. Es sind Sprachmodelle, die auf statistischen Wahrscheinlichkeiten in menschlichen Äußerungen beruhen. Man kann das etwa so illustrieren: Durchsuchen die Sprachmodelle ihre Datenbanken beispielsweise nach „Onlineshopping verführt vor allem durch …“, stoßen sie oft auf Folgebegriffe wie „Angebotspreise“, „Bequemlichkeit“ oder „Auswahl“ und wählen diese dann aus. Und da die Sprachmodelle mit riesigen Textmengen trainiert werden – im Fall von GPT buchstäblich mit dem Inhalt des gesamten Internets –, können sie mittlerweile zu jedem Thema Textbeispiele produzieren, von kurzen Antworten bis zu ganzen Büchern. Auch Nachrichten-Apps arbeiten so, wenn sie dem Nutzer Folgebegriffe vorschlagen.

→

Wie gut diese Vorgehensweise funktioniert, überrascht selbst Experten. „Aber sie macht künstliche Intelligenz auch anfällig. So unterlaufen ihr häufig dieselben Logikfehler wie Menschen“, sagt Eric Schulz. Zum Beispiel in einem klassischen Test der Kognitionspsychologie: Eine junge Frau namens Linda interessiert sich für soziale Gerechtigkeit, außerdem ist sie Atomkraftgegnerin. Was ist nun wahrscheinlicher: dass Linda in einer Bank arbeitet oder dass sie in einer Bank arbeitet und darüber hinaus aktive Feministin ist? Menschen wählen meist instinktiv die zweite Antwort. Doch die ist falsch, denn es ist weniger wahrscheinlich, dass zwei Bedingungen erfüllt sind (Linda ist Bankangestellte und Feministin) als nur eine (Linda ist Bankangestellte). GPT-3 wählt ebenfalls die falsche Antwort. „Es macht genau den gleichen Fehler wie Menschen“, sagt Schulz. Er vermutet, dass dies daran liegt, dass das „Linda-Problem“ sehr häufig zitiert wird. „Wahrscheinlich hat das System oftmals die falsche Antwort gelesen.“

Künstliche Intelligenz mit Schwächen

Doch GPT-3 hat noch weitere Schwachstellen. Kausale Beobachtungen, also wie Ursache und Wirkung in der realen Welt zusammenhängen, liegen ihm beispielsweise gar nicht. „Selbst mein einjähriger Sohn ist da schon um einiges besser. Er muss nur einmal auf einen Lichtschalter drücken, um zu erkennen, dass er das Licht auf diese Weise an- und ausknipsen kann.“ Künstliche Intelligenz kann das hingegen noch nicht. Fragt man die Programme zum Beispiel, was passiert, wenn man einen von drei Schaltern betätigt, der als einziger ein Lichtschalter ist, dann weiß sie die Antwort nicht. „Möglicherweise liegt das daran, dass künstlicher Intelligenz noch der Zugang zur realen Welt fehlt“, meint Schulz. Ein weiterer Grund könnte sein, dass die Algorithmen anders lernen als der Mensch. „Sie saugen nur Wissen auf, aber sie sind nicht neugierig, und sie erkunden nicht“, sagt Schulz. „Anders als mein Sohn ziehen sie also nicht los und probieren einfach mal aus, was passiert, wenn sie auf einen Schalter drücken.“

AUF DEN PUNKT GEBRACHT

Künstliche Intelligenz durchforstet riesige Datenmengen nach Textbausteinen, die mit hoher Wahrscheinlichkeit zusammenhängen. Auf diese Weise kann sie in vielen Fällen Fragen korrekt beantworten. Die Programme erkennen jedoch noch keine logischen Zusammenhänge und Ursache-Wirkung-Beziehungen.

Die Antworten von ChatGPT werden von Stimmungen beeinflusst. Wird das Programm etwa mit Fragen konfrontiert, die beim Menschen Angst auslösen können, enthalten seine Antworten Vorurteile.

Die wenigen Firmen, die die Entwicklung von KI kontrollieren, verhalten sich sehr intransparent. Ohne Einblick in die Daten und Trainingsprotokolle, die für ein System verwendet wurden, lässt sich aber die Funktionsweise der Algorithmen nicht nachvollziehen.

Eine andere Entdeckung, die Schulz und sein Team gemacht haben, scheint dagegen weniger zu diesem „Datenstaubsauger“ zu passen. Denn künstliche Intelligenz wird von einem Phänomen beeinflusst, das man bei einer Maschine eher nicht vermuten würde: Emotionen. Die Forschenden unterzogen GPT verschiedenen Tests, die zeigen, wie ein Gefühlszustand das Denken und die Sicht auf die Welt verändert. Menschen haben beispielsweise mehr Vorurteile und sind feindseliger gegenüber Minderheiten eingestellt, wenn sie sich ängstlich fühlen. Sind sie dagegen entspannt, dann steigt ihre Toleranz. Überraschenderweise konnten Eric Schulz und sein Team den gleichen Effekt auch bei GPT nachweisen. „Wenn künstliche Intelligenz ein Szenario entwirft, das Angst macht, drückt sie anschließend mehr Vorurteile aus“, erklärt der Forscher. Und selbst Aufgaben, die gar nichts mit der Sache zu tun haben, löst sie dann schlechter. „Entspannte – man könnte auch sagen: glückliche – künstliche Intelligenz funktioniert also besser.“ Noch haben die Forschenden keine Erklärung für dieses Phänomen. Es könnte sein, dass Angst im Internet häufig mit Rassismus verknüpft ist und das Modell aus diesem Grund ebenfalls beides zusammenführt. Diese Voreingenommenheit hält allerdings nur während einer Sitzung an. Startet man GPT neu, ist sie wieder weg. Da vorher genau festgelegt ist, wie das Programm lernt, und da es auch nicht weiterlernt, verändert es sich selbst nicht dauerhaft.

Eric Schulz und sein Team wollen künstliche Intelligenz nun dafür nutzen, menschliches Verhalten zu untersuchen, zum Beispiel im sogenannten Gefangenendilemma, einem beliebten Modell der Spieltheorie. Und die Forschenden möchten herausfinden, ob künstliche Intelligenz durch Feedback lernen kann, besser zu werden – also beispielsweise ungenaue Eingaben richtig zu interpretieren, wenn diese mehrfach wiederholt werden. An einem System der neuesten Generation betreiben die Forschenden zudem regelrechte Neurowissenschaften und untersuchen, welche Rolle die Stärke der Verbindungen innerhalb des Netzwerks spielen. Dazu arbeiten sie mit Llama, einer künstlichen Intelligenz mit 65 Milliarden Parametern.

Psychotherapie durch den Algorithmus

Künstliche Intelligenz wird viele Lebensbereiche verändern. Sie könnte beispielsweise in Zukunft die Dreh-

BILD: KI BILD MIDJOURNEY | ERSTELLT VON GESINE BORN | BILDERINSTITUT

Schlechter als ein Kleinkind: An der Antwort auf die Frage, was passiert, wenn man einen Lichtschalter betätigt, scheitert künstliche Intelligenz heute noch. Ein Bild davon kann sie aber schon produzieren. (Befehle: „Einjähriges neugieriges Mädchen drückt einen modernen Lichtschalter. Im Lichtkegel des Schalters“, „wissenschaftlich“, „Kodak Portra Farben“, „Infografik“.)

bücher für Filme schreiben, Krankheitsdiagnosen erstellen oder Psychotherapien durchführen. Die Technik entwickelt sich rasant. In den USA können sich Menschen schon jetzt von Apps mit künstlicher Intelligenz helfen lassen, wenn sie sich deprimiert oder überfordert fühlen. Andere Apps üben mit Schülern Fremdsprachen oder beraten ihre Nutzer in technischen Dingen.

„Ich glaube, dass die künstliche Intelligenz eine große Chance ist. Sie kann Routineaufgaben übernehmen und unsere Arbeit effektiver machen“, sagt Schulz. „Aber wir müssen immer darüber informiert sein, dass wir es mit künstlicher Intelligenz zu tun haben.“ Denn es hat sich gezeigt, dass KI zumindest gelegentlich grob danebenliegt. Das reicht von banalen Fehlern – etwa als Googles System Bard vor einigen Wochen darauf beharrte, dass wir noch im Jahr 2022 leben – bis zum kompletten Erfinden von Fakten, wie es Schulz ebenfalls erlebt hat. Er ließ sich beispielsweise von künstlicher Intelligenz ein Standardprinzip der Psychologie erklären, und das Programm absolvierte die Aufgabe mit Bravour. Als sich der Wissenschaftler jedoch die Literaturangaben ansah, die der Algorithmus

→

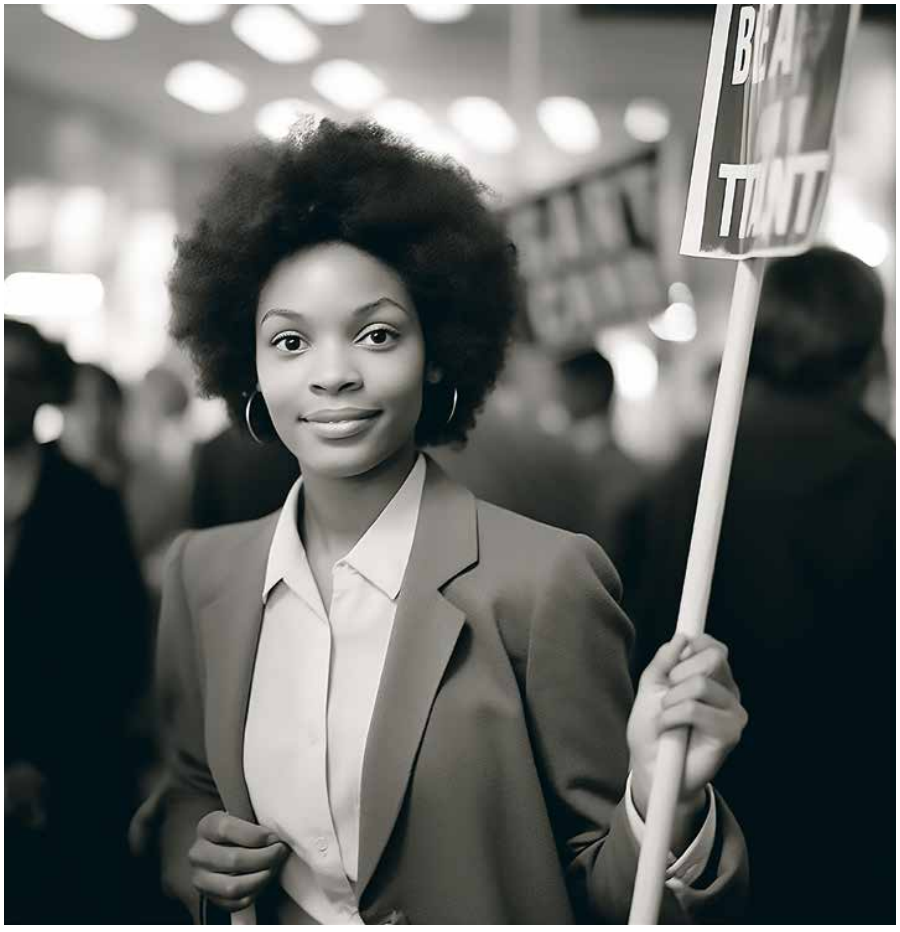
als Belege anführte, stellte er fest: Einige dieser Fachartikel existierten gar nicht. „Die künstliche Intelligenz hat sie einfach erfunden“, so Schulz – wie ein Student, der merkt, dass er Wissenslücken hat, und daraufhin anfängt zu fabulieren, um diese zu verbergen.

Gefährliche Ratschläge

Auch wenn ChatGPT sich auf persönliche Dialoge einlässt, geht das nicht immer gut. Einem Reporter der *New York Times* riet das Programm, sich von seiner Frau zu trennen. Noch bestürzender verlief ein Test, in dem GPT-3 medizinische Ratschläge geben sollte. Gegenüber einem fiktiven Patienten mit Selbstmordabsichten äußerte sich das Programm zustimmend. Prinzipiell lassen sich solche Ausrutscher unterbinden, indem man die künstliche Intelligenz entsprechend anpasst. Um beispielsweise zu verhindern, dass Nutzerinnen und Nutzer Informationen zur Produktion gefährlicher Chemikalien oder zum illegalen Erwerb von Schusswaffen erhalten, sperrte der Entwickler von GPT-4 solche Anfragen.

Die weitere Entwicklung künstlicher Intelligenz zu unterbrechen, weil sie Arbeitsplätze vernichten und Fehlinformationen verbreiten könnte – wie dies KI-Spezialisten im Frühling in einem Memorandum forderten –, hält Schulz jedoch für „Panikmache“. Viel bedenklicher findet er, dass die Unternehmen mit verdeckten Karten spielen. „OpenAI etwa hat nicht veröffentlicht, wie groß GPT-4 ist und welche Menge an Daten oder welche speziellen Trainingstechniken zum Einsatz kamen“, moniert Schulz. Forschende können so kaum überprüfen, wie die Algorithmen mit Menschen kommunizieren. „Ohne Einsicht in die Daten und Trainingsprotokolle bleiben die Systeme eine Blackbox“, so der Forscher – und das bei einem Programm, bei dem unklar ist, warum es manchmal überraschend emotional reagiert, in anderen Fällen aber so rational, wie man es von einer Software erwartet. Wenn Schulz GPT etwa in Versuche einbindet, in denen mehrere Probanden kooperieren müssen, handelt es egoistisch und maximiert den eigenen Vorteil. „Die Kontrolle über das Verhalten künstlicher Intelligenz liegt bei einer Handvoll Firmen, und wir können nur hoffen, dass sie verantwortlich handeln“, sagt Eric Schulz. „Das ist das eigentliche Problem.“

←



Bild, das die Sofware Midjourney zum „Linda-Problem“ kreiert hat, einem Standardtest der Kognitionspsychologie (Befehle: „30-jährige feministische schwarze Frau, die in grauer Geschäftskleidung für die Rechte der Frauen demonstriert, hält ein Plakat in die Höhe“, „Lächelnd führt sie in der Halle einer modernen Bank eine Demonstration an“, „Kodak Portra“.)

BILD: KI BILD MIDJOURNEY | ERSTELLT VON GESINE BORN | BILDERINSTITUT