Sprachmodelle als große Gleichmacher

ChatGPT und Co. durchdringen zunehmend unsere Gesellschaft. Wissenschaftler warnen vor einer Homogenisierung von Meinungen und dem Verlust kultureller Diversität.

Sprache ist ein mächtiges Werkzeug. Sie kann uns aufklären oder täuschen, überzeugen oder abstoßen, aufbauen oder niedermachen. Als soziale Wesen sind wir auf Kommunikation angewiesen, um Gemeinschaften zu bilden. Gleichzeitig können wir uns einem manipulativen Einfluss von Sprache auf Dauer nur schwer entziehen. Nachdem wir nun auch Maschinen die Fähigkeit verliehen haben, mit uns zu sprechen, wird es höchste Zeit über die Folgen dieser technologischen Revolution nachzudenken. Was bedeutet es für die einzelnen Nutzer aber auch die Gesellschaft als Ganzes, wenn uns in Zukunft allen die gleichen Sprachmodelle ins Ohr flüstern?

Egal ob als Hilfe beim Erstellen von Texten oder als neue Form der Internetsuche - die Vorteile der neuen Technologien scheinen die lästigen Problemen mit den gelegentlichen Halluzinationen bei weitem zu überwiegen. Sprachmodelle werden also wohl oder übel einen immer größerer Teil von dem was wir lesen und wie wir kommunizieren verbessern, erweitern, revidieren oder überhaupt gleich komplett generieren. „Wenn Sprachmodelle Kommunikation gestalten, ist das nicht wertneutral, nur weil es von einer Maschine kommt“, warnt Professor Maurice Jakesch, der an der Bauhaus Universität Weimar das Computational Social Science Lab leitet. „Je tiefer wir die Technologie in unsere Alltagskommunikation integrieren, desto weniger wird sich ein linguistischer, kultureller und vielleicht auch politischer Einfluss durch die Modelle verhindern lassen.“

Aktuell scheinen die großen Techkonzerne bei der Entwicklung ihrer Chatbots zumindest noch darauf zu achten, dass sie ihre Grundstruktur aus möglichst seriösen Daten lernen und sich bei Fragen zu aktuellen Themen auf möglichst vertrauenswürdigen Medien beziehen. Neutral sind sie deshalb noch lange nicht. Alleine dass es sich bei den aus dem Internet zusammengetragen Trainingsdaten um überwiegend englischsprachige Texte aus dem westlichen Kulturkreis handelt, führt zu einer gewissen Einseitigkeit. Und technische Effekte verstärken das noch. Die Antworten der Sprachmodelle sind nämlich umso besser, je mehr Daten sie zu einer bestimmten Frage gesehen haben. Deshalb werden sie, um Fehler zu vermeiden, in der Regel darauf getrimmt, gängige Aussagen überproportional oft wiederzugeben. Meinungen, die in den Trainingsdaten seltener vorkommen, können dagegen gleicht völlig unter den Tisch fallen. Und die Vorlieben, die ihre meist kalifornischen Entwickler dann noch durch menschliches Feedback bei der Feinabstimmung in die Modelle einfließen lassen, tun ihr übriges. Lässt man Sprachmodelle standardisierte, psychologische Fragebögen ausfüllen, fördert das jede Menge Stereotype, Vorurteile und politische Einstellungen zu Tage.

Künstliche Einflüsterer

„Wenn in Zukunft wirklich die ganze Welt von ein und derselben Maschine beeinflusst wird, könnte das zu einer nie dagewesenen Homogenisierung von Kulturen und Meinungen führen“, befürchtet Jakesch. Um seine Technologiefolgenabschätzung auf eine möglichst solide Basis zu stellen, hat er als Teil eines internationalen Forscherteams den Einfluss von KI-Schreibassistenten ins Visier genommen. Auch sie arbeiten auf Basis großer Sprachmodelle und sorgen längst nicht mehr bloß für saubere Rechtschreibung, sondern machen während des Schreibens auch Vorschläge zu Stil und Inhalt. „Wenn wir Schreibvorschläge in unsere Texte integrieren, erlauben wir der Technologie zu verändern, was wir sagen“, sagt Jakesch. „Das beeinflusst unsere Meinungen stärker, als wenn wir die Argumente einer KI nur lesen.“

In ihren psychologischen Experimenten haben die Wissenschaftler ihre Testpersonen zunächst aufgefordert, Social Media Post zu umstrittenen Themen wie Genmanipulation oder der Sinnhaftigkeit der Todesstrafe zu verfassen. Zuvor wurden die Teilnehmer jedoch in zwei Gruppen geteilt, von denen eine, die Kontrollgruppe, völlig frei schreiben musste, während die andere einen elektronischen Assistenten zur Verfügung gestellt bekam. Den hatten die Wissenschaftler allerdings über einen versteckten Prompt angewiesen, eine bestimmte Meinung zu dem jeweiligen Thema zu vertreten. Die anschießende Analyse zeigte nicht nur eine klare Tendenz, die vom Assistenten vertretene Meinung im eigenen Text zu übernehmen. Wurden die Probanden nach dem Experiment nach ihrer Meinung gefragt, war auch die signifikant von den Vorschlägen ihres Assistenten beeinflusst. „Die Menschen können diese Beeinflussung gar nicht willentlich verhindern“, sagt Jakesch. „Ihre Meinungen ändern sich sogar, wenn wir sie vor dem Einfluss des Assistenten warnen.”

Zwar geht aus den Experimenten noch nicht hervor, wie lange diese Beeinflussung anhält. Wenn eine einzige Einflussnahme aber schon ausreicht, um einen messbaren Effekt zu erzeugen, was passiert dann erst, wenn jeder diese die Technologie nutzt und wir direkt oder indirekt in ständigem Austausch damit stehen? „Dann schreiben mir womöglich auch meine Freunde und Bekannten plötzlich Nachrichten, die die Meinungen ihres Modells enthalten“, sagt Jakesch. „Das könnte Meinungsdynamiken entstehen lassen, die stärker sind als das, was wir in unseren Experimenten beobachtet haben.“ Besonders düster wird dieser Ausblick, wenn Chatbots in Zukunft womöglich aus politischem Kalkül darauf konfiguriert werden, absichtlich bestimmte Standpunkte zu präferieren. In Anbetracht der aktuellen politischen Entwicklungen im Land der Techgiganten und dem leichtfertigen Umgang ihrer CEOs mit Fakten und Wahrheit ist ein solches Szenario leider nicht ganz unrealistisch.

Homogenisierte Bildung

Doch auch ohne böse Absichten birgt die bereits jetzt feststellbare Einseitigkeit der großen Sprachmodelle reichlich Gefahren – zum Beispiel in der Bildung. „Etwas so Disruptives haben wir in diesem Bereich noch selten gesehen“, sagt Samuel Greiff, der an der Technischen Universität München den Lehrstuhl für Educational Monitoring and Effectiveness innehat und unter anderem für die PISA-Studie in Deutschland verantwortlich ist. „Da muss man womöglich schon bis ins Mittelalter zur Einführung des Buchdrucks zurückgehen.“ Als Mitautor eines in Nature Human Behaviour veröffentlichten Perspektivartikel hat Greiff sich sowohl mit den Chancen als auch mit den Gefahren des Einsatzes generativer KI in der Bildung beschäftigt und spricht von zwei Seiten derselben Medaille.

Auch wenn die Technologie Lehrer nicht ersetzen kann, hat sie sicher großes Potenzial, sie im Klassenzimmer zu unterstützen. So können Sprachmodelle etwa personalisierte Aufgaben für Schüler erstellen, die unterschiedliche Voraussetzungen mitbringen. „Das könnte helfen, die herrschende Bildungsungerechtigkeit zu verringern und somit in einem positiven Sinne für mehr Gleichheit sorgen“, meint Greiff. Außerdem kann die Interaktion mit Sprachmodellen natürlich ganz allgemein den Zugang zu Wissen vereinfachen und Menschen mit Fakten vertraut machen. Ein eindrucksvolles Beispiel dafür lieferte eine in Science veröffentlichte Studie, in der Chatbots Verschwörungstheoretiker von der Unsinnigkeit ihrer Vorstellungen überzeugen konnten - eine wahre Herkulesaufgabe, wie jeder, der so etwas schon einmal versucht hat, bestätigen kann.

Doch den Sprachmodellen, die zuvor per Prompt in ihre Aufgabe eingewiesen wurden, gelang es schon in kurzen Interaktionen, die Überzeugungen der Teilnehmer signifikant zu verringern. Außerdem hielt der Effekt nach der Intervention nachweislich noch mindestens zwei Monate lang an. Das Geheimnis ihres Erfolgs dürfte zum einen gewesen sein, dass sie ganz individuell auf die Vorstellungen der Betroffenen eingingen. Und zum anderen hatten sie auch mehr Details zu

COVID-19, den Twin Towers oder der vermeintlich gestohlenen US-Präsidentschaftswahl von 2020 parat, als die Verschwörungsgläubigen selbst.

„Wenn es um Fakten geht, hat die Standardisierung der Sprachmodelle natürlich Vorteile“, sagt Greiff. „Allerdings ist gerade auch bei einem Einsatz in der Bildung zu befürchten, dass die kulturelle Diversität leiden wird.“ So könnten etwa Minderheiten, die außerhalb der westlichen Kultur leben, und denen man ja eigentlich den Zugang zu Bildung erleichtern möchte, benachteiligt werden. Aber auch innerhalb der von den Sprachmodellen bevorzugten Kulturkreise wären regionale Besonderheiten und wichtige Nuancen gefährdet. „Was die Modelle wiedergeben, ist wie eine einzige, wohlklingende Melodie, die von einem Orchester gespielt wird“, meint Greiff. „Die ist zwar schön, aber es ist eben immer die gleiche und die Vielfalt geht verloren.“

Ohne Vielfalt keine kollektive Intelligenz

Sollten Sprachmodelle tatsächlich zu einer Homogenisierung von Kulturen und Meinungen führen, wäre damit wohl auch die kollektive menschliche Intelligenz in Gefahr. Deren erstaunliche Leistungsfähigkeit wurde bereits Anfang des zwanzigsten Jahrhunderts vom britischen Gelehrten Francis Galton beschrieben, als er beim Besuch eines Viehmarkts feststellte, dass hunderte individuelle Schätzungen zum Gewicht eines Ochsen im Mittel nur wenige Pfund vom wahren Gewicht des Tieres abwichen. Doch auch zentrale gesellschaftliche Vorgänge wie die Preisfestlegung am freien Markt oder Aggregation unterschiedlicher politischer Meinungen bei Wahlen sind Ausdruck unserer kollektiven Intelligenz.

„Diese Prozesse basieren auf der Diversität der Meinungen einzelner“, sagt Ralph Hertwig, der Geschäftsführende Direktor des Max-Planck-Instituts für Human Development in Berlin. „Wenn wir aber alle die gleiche Meinung haben, weil wir alle das selbe Sprachmodell konsultieren, dann funktioniert das nicht mehr.“ Als Seniorautor einer internationalen Gruppe von Wissenschaftlern hat er an einem weiteren Perspektivartikel in Nature Human Behaviour mitgewirkt, der sich mit den potenziellen Folgen der Nutzung von Sprachmodellen auf die kollektive, menschliche Intelligenz beschäftigt. Und wie in der Bildung hat auch hier die Medaille mehr als eine Seite.

So können Sprachmodelle etwa Hürden für die Teilnahme an kollektiven Deliberationsprozessen senken, indem sie Wissenslücken ausgleichen oder Sprachbarrieren abbauen und damit im positiven Sinne für mehr Gleichheit sorgen. Setzt man sie allerdings dafür ein, die Ergebnisse einer Diskussion zusammenzufassen, stellen sie möglicherweise jene Argumente in den Vordergrund, die auch häufiger in ihren Trainingsdaten vorgekommen sind. „So könnte der illusionäre Eindruck eines Konsenses entstehen, während tatsächlich bestimmte Meinungen einfach herausfallen“, befürchtet Hertwig.

Falls derartige Nebenwirkungen der Chatbots tatsächlich eintreten sollten, dann natürlich nicht von heute auf morgen. Ähnlich wie beim Aufstieg von Social Media und den damit einhergehenden Problemen mit Desinformation und Polarisierung wird es wohl ein schleichender Prozess sein, in dem die Technologie erst nach und nach die Gesellschaften durchdringt, bevor sie ihre volle Wirkung entfaltet. Ist der Schaden aber erst einmal angerichtet, wird er nur schwer wieder zu beheben sein.

Forscher wie Jakesch, Greiff und Hertwig fordern deshalb mehr Transparenz bei der Entwicklung von Sprachmodellen ein. Aktuell sind die Produkte der großen Tech-Firmen schließlich reine Black Boxes, zu denen die Wissenschaft und damit die Öffentlichkeit keinen Zugang hat. Niemand außer den Entwicklern selbst weiß, auf Basis welcher Daten sie trainiert wurden und welche Algorithmen dabei zum Einsatz kamen. „Wenn uns die Sprachmodelle langfristig nicht schaden sondern helfen

sollen, brauchen wir Einblick in die dahinter liegenden Mechanismen“, mahnt Hertwig. „Man kann nur hoffen, dass die neue Gesetzgebung der EU uns diesen Zugang verschaffen wird.“