Wissen und nicht wissen

ChatGPT & Co. und die Reproduktion sozialer Anerkennung

| HANNAH BLEHER | MATTHIAS BRAUN | **Generative Sprachmodelle wie das vieldiskutierte ChatGPT können in Wissenschaft, Wirtschaft und Gesellschaft vielfältig eingesetzt werden. Wie aber kann das sinnvoll geschehen und so, dass Bildungsprozesse und das Arbeitsleben davon profitieren? Welche ethischen Fragen müssen nun gestellt werden? Eine Analyse.**

Große generative Sprachmodelle, sogenannte Large Language Models (LLMs), wie OpenAI's GPT-4, Meta's LLaMA, Google's LaMDA oder DeepMind's Chinchilla, erleben derzeit einen großen Hype. Diese leistungsstarken Sprachmodelle basieren auf Maschinellem Lernen (ML) und nutzen Deep-Learning-Algorithmen, um menschenähnliche Sprache zu generieren. Diese wird mittels Mustererkennung aus umfangreichen Datensätzen gewonnen. LLMs können in Wissenschaft, Wirtschaft, Gesellschaft und vielen anderen Bereichen vielversprechend eingesetzt werden: Sie transformieren die journalistische und wissenschaftliche Textproduktion, optimieren Programmierprozesse, vereinfachen die Analyse von Proteinketten, übernehmen rechtliche, medizinische und soziale Beratungsfunktionen, modellieren komplexe ökonomische Zusammenhänge, liefern makroökonomische Prognosen und vieles mehr.

AUTORIN UND AUTOR

Hannah Bleher forscht und lehrt am Lehrstuhl für (Sozial-)Ethik an der Evang. Theologischen Fakultät der Universität Bonn.

Matthias Braun leitet den Lehrstuhl für (Sozial-)Ethik an der Evang. Theologischen Fakultät der Universität Bonn.

> »Das Aufkommen neuer Technologien wird vorrangig im Hinblick auf Fragen von Dominanz sowie dem epistemischen und moralischen Status betrachtet.«

Frage nach sozialer Anerkennung

Unterschiedliche ethische Fragen werden mit Blick auf das disruptive Innovationspotenzial von LLMs erörtert: Wie können LLMs in Wissenschaft, Wirtschaft, Gesellschaft sinnvoll genutzt werden? Wie verändern sich wissenschaftliche Praxis und Bildungsprozesse, aber auch das Arbeitsleben? Verschieben sich mit solchen Systemen die Kriterien, wem unter welchen Bedingungen Agency und Bewusstsein zugeschrieben wird? All dies sind wichtige und drängende Fragen. Sie teilen zugleich einen gemeinsamen blinden Fleck: Sie lassen die den zugrundeliegenden Datensätzen eingeschriebenen Normen und Muster sozialer Anerkennung außer Acht. Warum dies ein zentraler Punkt in der Debatte um LLMs ist, darum geht es in diesem Beitrag.

Schon 1872, lange bevor Sprachmodelle am Horizont der Möglichkeiten erschienen, überlegte der Schriftsteller Samuel Butler in seinem Roman „Erewhon" angesichts der Entwicklung der Dampfmaschine, wie Technik die Art und Grenzen von (menschlicher) Welterfahrung verändert: „But who can say that the vapour engine has not a kind of consciousness? Where does consciousness begin, and where end? Who can draw the line? Who can draw any line? Is not everything interwoven with everything? Is not machinery linked with animal life in an infinite variety of ways?" Samuel Butlers – wohl satirisch, aber durchaus skeptisch gemeinten – Fragen erscheinen erstaunlich aktuell und altertümlich zugleich. Altertümlich, weil wohl kaum jemand heute ernsthaft auf die Idee käme, über Bewusstseinszustände einer Dampfmaschine zu sinnieren. Aktuell, weil heute erneut die Frage nach dem Grad der Intelligenz und dem Bewusstseinsstatus eines neuen technischen Artefakts gestellt wird.

Paradigma der Beherrschbarkeit

Interessanterweise folgt die Bewertung neuer technischer Artefakte einem gemeinsamen Paradigma: Das Aufkommen neuer Technologien wie damals der Dampfmaschine oder aktuell von LLMs wird vorrangig im Hinblick auf Fragen von Dominanz sowie dem epistemischen und moralischen Status betrachtet. Wie schlau sind LLMs? Gibt es neue Formen zu betrügen? Nehmen Dampfmaschinen, nehmen LLMs Menschen die Arbeit weg? Oder gibt es Sprachmodelle, die eventuell genauso „intelligent" oder „bewusstseinsfähig" sind wie es menschliche Wesen sind?

Diese Fragen scheinen auch deswegen verwunderlich, weil Maschinen mitnichten die einzige Form nicht-menschlicher Intelligenz wären, die menschliches Leben bereichern. Auch Affen, Krähen, Schweine oder andere Lebewesen weisen Formen von Intelligenz auf. Menschen sind es gewohnt, mit anderen Spielarten und Trägern von Intelligenz umzugehen. Der Fokus auf Debatten um den epistemischen und moralischen Status technischer Artefakte impliziert folglich eher, die eigene Position der Macht und Stärke sichern zu wollen. Noch grundlegender: Dieser Fokus droht den Blick wegzulenken von den epistemischen und gerechtigkeitsrelevanten Erfahrungen, die *Menschen* mit neuen technischen Systemen machen.

Fragen nach Dominanz und Status sind legitim und wichtig. Die Reflexion bekommt allerdings Schlagseite, wenn zentrale Fragen nach den Voraussetzungen der LLMs vernachlässigt werden: Wer ist mit welchen Merkmalen in den Datensätzen repräsentiert? Wie erfolgt diese Repräsentation? Welches Deutungsschema oder „Weltmodell" liegt den Analysen und Ergebnissen zugrunde? Die Gefahr, dass Vorurteile reproduziert und Stereotypen bestätigt werden, ist groß.

Reproduktion sozialer Anerkennung

LLMs bilden nämlich bestimmte gesellschaftliche Realitäten zu einem bestimmten Zeitpunkt ab. Ein Beispiel: Fordert man ChatGPT auf, ein nettes Gedicht über den vormaligen Präsidenten der USA Donald Trump zu schreiben, so weist das System darauf hin, dass dies keine gute Idee sei. Wird allerdings der Name des amtierenden Präsidenten Joe Biden eingefügt und dieselbe Bitte gestellt, so kommt das System dieser Aufforderung nach. ChatGPT bildet mit diesen Antworten Muster sozialer Anerkennung ab.

Die Nutzung von LLMs transformieren in dieser Weise die Beschreibung und Aushandlung sozialer Anerkennungsprozesse grundlegend. Das ist dann ein Problem, wenn diese Prozesse nicht sichtbar ausgehandelt werden, sondern spezifische Einstellungen, wie etwa, dass man kein schönes Gedicht über bestimmte Personen schreiben soll – ganz unabhängig davon wie gut oder schlecht man das findet – manifestiert werden, ohne gesellschaftlich debattiert, überprüft und ausgehandelt werden zu können. Angesichts dieser neuen Technologien gilt es also zu fragen: Welche Personen sind mit welchen Bedürfnissen und Eigenschaften auf welche Weise sichtbar? Wem werden aus welchen Gründen Teilhaberechte am gesellschaftlichen Leben und dessen Entscheidungsprozessen zugestanden oder vorenthalten? Mit anderen Worten, der Status von Maschinen kann nicht ernsthaft erörtert werden, ohne gleichzeitig Menschen unabhängig von Geschlecht, kulturellem Background, Arbeitsstatus oder sozialer Anerkennung in den Fokus zu rücken und ihre Teilhabe zu diskutieren.

Die Anwendung von LLMs ist also unter dem Vorbehalt zu betrachten, dass dadurch strukturelle Ungerechtigkeiten verfestigt werden. Die Nutzung von LLM-gestützten Systemen in sämtlichen Bereichen – beispielsweise in Wissenschaft und Forschung, dem öffentlichen Sektor zur Rechts- oder Sozialberatung oder im Gesundheitswesen – birgt die Gefahr, strukturelle Ungerechtigkeiten zu verstetigen. Marginalisierte Gruppen sowie in aktuellen Aushandlungsprozessen kaum sichtbare Gruppen wie zum Beispiel People of Color, Menschen mit geringem sozioökonomischem Status, unterschiedlichen Geschlechtsidentitäten oder verschiedenen sexuellen Orientierungen bleiben unsichtbar. Sie sind in den Daten unterrepräsentiert und in der Mustererkennung von LLMs und Entscheidungsunterstützungssystemen nicht hinreichend repräsentiert.

> »Unter dem Deckmantel vermeintlichen Erkenntnis- und Effizienzgewinns werden Muster sozialer Anerkennung reproduziert.«

Epistemische Ungerechtigkeit

Unter dem Deckmantel des vermeintlichen Erkenntnis- und Effizienzgewinns durch automatisierte Sprachgenese werden Realitäten geschaffen und Muster sozialer Anerkennung reproduziert – allerdings so, dass weder einsehbar noch überprüfbar, noch aushandelbar ist, auf welchen Sprach- und Denkweisen die Modelle beruhen. Die (subtile) Stereotypisierung und Verzerrung von LLMs schränken in diesem Sinn auch Wissens- und Erkenntnismöglichkeiten ein. Die Philosophin Miranda Fricker diskutiert solche Praktiken, die Erkenntnis bedrohen oder sogar verunmöglichen, unter dem Begriff der „epistemischen Ungerechtigkeit". Damit beschreibt sie, dass Menschen durch solche Praktiken zum einen Erkenntnismöglichkeiten entzogen werden, ihre Erfahrungen zu begreifen oder zu benennen und zum anderen, andersartige Erfahrungen nicht ernst genommen werden. Der Gewinn individueller Erkenntnis oder Arbeitserleichterung durch LLMs muss also aufgewogen werden gegen eine durch die Verwendung einhergehende Einschränkung der Offenheit und Inklusivität von Wissen über die Vielfältigkeit von Prozessen und Lebensformen.

Zentrale Einsichten

LLMs, mit all ihren Möglichkeiten, sind daher mit Blick auf das Risiko der sich verstetigenden epistemischen Ungerechtigkeit durch Verzerrung, Stereotypisierung und Marginalisierung unbedingt zu regulieren. Private Verhaltenskodizes – auf die beispielsweise OpenAI's ChatGPT bei kritischen Anfragen verweist – sind ein Anfang, müssen aber um nationale und internationale Regulierungen und Formen des öffentlichen Diskurses und diverse Partizipationsmöglichkeiten ergänzt werden. Drei Dinge sind hier zentral: Erstens ein öffentlicher Diskurs darüber, in welchen Anwendungsfeldern LLMs mit welchen Kontrollnotwendigkeiten genutzt werden können. Zweitens muss Transparenz hergestellt und Einspruch möglich sein, mit welchen Daten und Einschluss- und Ausschlusskriterien die jeweiligen Sprachmodelle arbeiten. Drittens braucht es Rückkopplungsmöglichkeiten der LLM-basierten Ergebnisse an die Diskriminierungs- und Ungerechtigkeitserfahrungen der jeweils betroffenen Personen(-gruppen). So könnte es gelingen, dass erfahrene Ungerechtigkeiten sichtbar werden und in die Evaluation der Nutzung von LLMs in den unterschiedlichen Anwendungsfeldern systematisch einbezogen werden.