Paola Lopez

ChatGPT und der Unterschied zwischen Form und Inhalt

Seit seiner Präsentation bekommt ChatGPT viel Aufmerksamkeit. Einige argumentieren, das Erstellen von Text werde positiv revolutioniert – andere fürchten eine Erosion verschiedenster textbasierter Institutionen wie etwa Zeitungen oder Beurteilungsmodi von Bildungsinstitutionen. Die meisten sind sich jedenfalls einig: ChatGPT und ähnliche Sprachmodelle sind,[1] weil besonders gut, etwas Großes. Gleichzeitig schwirren Beispiele durch das Internet, die zeigen, wie schlecht, falsch und unsinnig ein von ChatGPT produzierter Text sein kann. Diese Beurteilungsskala ist jedoch eindimensional – die Frage, »wie gut« ChatGPT funktioniert, geht an wesentlichen Punkten vorbei. Wie etwa: Was bedeutet »gut« in diesem Kontext? Was folgt daraus, dass ChatGPT »gut« ist? Und die wichtigste Frage: Was wollen wir eigentlich von Texten?[2]

Übersehen wird dabei gerne, dass die technischen Fähigkeiten von ChatGPT in ganz wesentlicher Weise beschränkt sind. Beschränkt meint hier eine systematische Einschränkung: Manche Dinge kann ChatGPT und manche nicht. Und welche es kann und welche nicht, hängt mit seinen mathematischen Eigenschaften zusammen. Auch ein Material wie etwa Holz eröffnet für seinen Gebrauch bestimmte Möglichkeitsräume.[3] Aus Holz kann man eine Menge herstellen: Stühle, Zahnbürsten, Papier, aber der Möglichkeitsraum, der durch die Materialität von Holz abgesteckt wird, ist natürlich beschränkt. Wenn wir einen hölzernen Tisch bauen und von diesem Tisch erwarten, dass er Elektrizität leitet, dann werden wir enttäuscht werden. Holz hat bestimmte Charakteristika, die es uns ermöglichen, abzuschätzen, für welche Zwecke wir es sinnvollerweise verwenden können und für welche nicht.

1 Es gibt mittlerweile immer mehr Nachfolge- und Konkurrenzmodelle. Ich beschränke mich hier auf ChatGPT.

2 In diesem Essay befasse ich mich mit Text als Output von ChatGPT. Es gibt zahlreiche andere Anwendungen, in denen etwa Grafiken erstellt und Teile von Programmierprojekten automatisiert werden.

3 Auf die Frage, ob Künstliche Intelligenz eine Methode, ein Material, ein System, ein Paradigma, ein Vermittlungsinstrument, eine Kulturtechnik, ein Werkzeug, etc. ist, wird hier nicht weiter eingegangen. In meiner Arbeit ist es mir wichtig, einen möglichst entzauberten Zugang zu KI zu haben, also vergleiche ich KI an dieser Stelle mit Holz.

Ähnlich ist es mit mathematischen Werkzeugen: Sie haben in ihnen angelegte mathematische Charakteristika. Manche Dinge werden durch diese Charakteristika, sozusagen durch die »mathematische Materialität«, ermöglicht und manche nicht. In diesem Sinn sind die folgenden Ausführungen zu verstehen: Was macht die Mathematik unter, hinter und in ChatGPT möglich und was nicht?

Reinforcement Learning from Human Feedback

Sprachmodelle werden auf riesigen Textdatensätzen trainiert. Das funktioniert umso besser, je mehr und je vielfältigere Texte sich in den zugrundeliegenden Trainingsdaten befinden. Im Jahr 2020 wurde von OpenAI das Sprachmodell GPT-3 vorgestellt, das viel mediale Aufmerksamkeit erregte, unter anderem durch Artikel wie *A robot wrote this entire article. Are you scared yet, human?* im *Guardian*.[4]

Während des Trainingsprozesses wird ein Modell von Sprache erstellt. Dieses Modell wird verwendet, um Texte automatisiert zu simulieren – diese simulierten Texte sind der Output von ChatGPT. An dieser Stelle ist es wichtig, zu klären, was ein Modell ist. Ein Beispiel sind mathematische Gleichungen, die die Fortbewegung einer Welle in Wasser simulieren sollen (übrigens ein sehr schwieriges Unterfangen):[5] Es werden mögliche Parameter eingebaut, die den Prozess determinieren und beschreiben, und schließlich drückt man mit gewissen Startbedingungen sozusagen auf »Play«, lässt das Modell laufen, und das Modell simuliert die Fortbewegung der Welle bei verschiedenen Gegebenheiten wie Tiefenänderung, variierender Bodenreibung und so weiter. Am Ende kann man vielleicht abschätzen, wann eine Welle die Küste erreicht, ob und wie sie bricht oder wie sie sich weiter verhält: Wenn eine Welle mit gegebenen Parametern zu einem bestimmten Zeitpunkt an einer Stelle ist, wo ist sie dann zwei Sekunden später, und wie sieht sie aus? Diese Prognosen werden durch Modelle ermöglicht. Das Modell ist die Abstraktion einer idealtypischen Welle. Es funktioniert unabhängig

4 In: *Guardian* vom 8. September 2020 (www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3).

5 Ein wesentlicher Unterschied ist dabei, dass die Modellierung von Wasserwellen auf Gleichungen beruht, die zugrundeliegende Gesetzmäßigkeiten modellieren. Hinter den datenbasierten Methoden der KI steht – in den meisten Fällen – keine Theorie über die dahinterliegenden Gesetzmäßigkeiten. Diese werden aus den Trainingsdaten destilliert. Natürlich kann man die zwei Methoden der Modellierung auch kombinieren. Bei Sprachmodellen handelt es sich aber um eine rein datenbasierte Methode.

von einer konkreten Welle und kann bestenfalls in weiterer Folge in verschiedensten Kontexten eingesetzt werden.

Ein Modell soll also eine vereinfachte Version des Echten sein.[6] ChatGPT beinhaltet ein Modell von Sprache an sich, genauer: Text an sich. Die Frage, was Text-an-sich sein soll, kann unendlich diskutiert werden – wurde bei der Entwicklung von ChatGPT aber bereits implizit festgelegt: Text ist laut den Entwicklerinnen und Entwicklern von ChatGPT das, was sich in den Trainingsdaten befindet: dem Common-Crawl-Datensatz[7] sowie in vielen Büchern und großen Teilen von Wikipedia. Das ist immer so bei Methoden der Künstlichen Intelligenz: Alles ist, wie es ist, weil es in den zugrundeliegenden Daten so aufscheint – in diesem Fall Text.

Dabei ist es nicht so, dass ChatGPT bei jeder Erstellung eines Outputs den gesamten Trainingsdatensatz neu »durchschaut«. Es ist auch nicht so, dass bei jeder ChatGPT-Eingabe etwa das Internet nach passenden Informationen durchsucht wird.[8] Das ist, wie bei der Modellierung von Wasserwellen, alles bereits geschehen. Die statistisch für relevant befundene Information über Sprache aus dem Trainingsdatensatz ist schon in destillierter Form im Sprachmodell enthalten – sie *ist* in gewisser Weise das Sprachmodell. Das Modell, sobald fertig trainiert, existiert unabhängig vom Trainingsdatensatz, so wie ein Modell einer Welle unabhängig von konkreten Wellen existiert. Nur dann, wenn die Entwickler bei OpenAI beschließen, dass weiter- oder neu trainiert wird und neue Daten inkludiert werden – dann verändern sich die Parameter des Modells und die erzeugten Texte in weiterer Folge. Das ist auch der Grund, warum man bei der Anwendung von ChatGPT darauf hingewiesen wird, das Programm habe »limited knowledge of world and events after 2021« – die Daten kennen die Welt nach 2021 nicht.

6 Das ist natürlich gar nicht so einfach und alles andere als unbestritten. Jede Erstellung eines Modells beinhaltet unzählige menschliche Entscheidungen, die den Modellierungsprozess prägen. Das ist schon bei der Modellierung von Naturphänomenen so und umso mehr, wenn man soziale Phänomene oder Kulturtechniken wie Sprache modellieren möchte. Dennoch wäre es zu kurz gegriffen, zu argumentieren, dass jedes Modell vollends subjektiv ist, ausschließlich die Werte und Vorannahmen der Modelliererinnen beinhaltet und damit unbrauchbar ist. Die Frage, wie man ein Modell baut und was man mit einem Modell macht, trifft in den Kern von Aushandlungen und Auseinandersetzungen zwischen Wissenschaft und Politik. Vgl. Florian Eyert, *Das Mathematische ist politisch*. In: *WZB Mitteilungen*, Nr. 168, Juni 2020 (bibliothek.wzb.eu/artikel/2020/f-23105.pdf).

7 Common Crawl besteht aus weiten Teilen des Texts, der im Internet auffindbar ist, und beinhaltet mehr als eine Billion Wörter (commoncrawl.org/the-data/).

8 Es gibt Modelle, die ein Sprachmodell mit einer automatisierten Internetsuche koppeln.

Das zugrundeliegende Sprachmodell produziert jedes Wort (genauer: jeden Wortteil) einzeln und iterativ, es tritt nach jedem Wort einen Schritt zurück, blickt auf den bisher produzierten Satz und eruiert dann aufs Neue, welches ein passendes nächstes Wort sein könnte. Ein solches Sprachmodell enthält also eine iterative Wortteilprognose. Alle bisherigen Ausführungen treffen auch auf die Vorgängermodelle GPT (2018), GPT-2 (2019) und GPT-3 (2020) zu. Wie genau die Prognose funktioniert und wie es möglich ist, semantisch sinnvoll wirkende Sätze zu produzieren, werde ich hier auslassen. Ich beschränke mich auf die Neuerungen, die ChatGPT von seinen Vorgängermodellen unterscheiden. Der Sprung zwischen GPT-3 und ChatGPT besteht in der Kopplung eines Sprachmodells mit einem zusätzlichen Feedbackmodell: Mit einem bestimmten Verfahren wurde ein Modell gebaut, das menschliches Feedback automatisieren soll. Das Verfahren heißt »Reinforcement Learning from Human Feedback« (RLHF).

Ein gewisser Feedbackmechanismus ist im Machine Learning zwar immer eingebaut, da das Training Iterationsschleifen enthält, die menschliches Feedback einarbeiten. Hier ist es aber so, dass ein ganzes Modell entwickelt wurde, dessen Aufgabe es ist, zu einem gegebenen vom Sprachmodell produzierten Text ein Gut/Schlecht-Feedback zu geben. Das Training dazu funktionierte wie folgt: Das Feedbackmodell soll erkennen können, ob ein gegebener Text gut ist. Ein Team von vierzig eigens dafür beschäftigten Mitarbeitern hat, unter anderem, verschiedene Textausgaben des Sprachmodells für eine bestimmte Anfrage von gut nach schlecht geordnet. Dieses Ordnen wurde durch das Feedbackmodell automatisiert: Es »lernte« im Training, zu einem gegebenen Input verschiedene zufällig generierte Outputs von gut nach schlecht zu ordnen.

Der technisch geschickte Kniff bestand darin, das Feedbackmodell an die Textprognose zu koppeln. So produziert das Sprachmodell verschiedene Texte zu einer gegebenen Anfrage, und das Feedbackmodell wählt aus, welcher der Beste ist. Das inkludiert auch die Richtigkeit von produzierten Inhalten und die Vermeidung von schädlichen und anstößigen Outputs. Dieses Feedbackmodell wurde dazu benutzt, das Textprognosemodell weiter zu trainieren, im »fine tuning«, und sollte eine Lösung für das Problem sein, dass Training viele Daten erfordert, es aber teuer ist, Menschen dafür zu bezahlen, Texte zu bewerten: Man bezahlte einige Menschen und automatisierte den Bewertungsprozess.

Es ist also nicht so, dass beim tatsächlichen Einsatz eine Anfrage an ChatGPT gestellt wird, das Prognosemodell live zehn Antworten generiert, die live durch das Feedbackmodell geordnet werden, so dass dann die beste Antwort als Output generiert wird. Das ist alles im Vorfeld geschehen. Es

wurde also durch menschliche Eingaben ein Feedbackmodell trainiert, das in weiterer Folge dazu benutzt wurde, das Sprachmodell zu trainieren. Das ist das Erfolgsgeheimnis von ChatGPT. Es geht nicht mehr nur darum, ein wahrscheinliches nächstes Wort zu prognostizieren, sondern darum, ein wahrscheinliches nächstes Wort bei gleichzeitiger Optimierung des eingebauten »human feedback« zu erzeugen.

Die vierzig Menschen, die diese Bewertungen vornahmen, sind laut der Begleitpublikation zu ChatGPT größtenteils unter 35, zu gleichen Teilen Frauen und Männer und haben zum größten Teil südostasiatische oder US-amerikanische Staatsangehörigkeit.[9] In einer Umfrage, die OpenAI selbst durchgeführt hat, gaben fast 90 Prozent von ihnen an, sie seien ihrer Ansicht nach fair entlohnt worden.

Opazität versus Kontrolle

Die Soziologin und Informatikerin Jenna Burrell unterscheidet verschiedene Arten von Opazität, also Arten der Black-Box-Haftigkeit von Machine-Learning-Systemen:[10] Zum einen gibt es die absichtliche, institutionelle Opazität. Unternehmen veröffentlichen nicht die Informationen, die notwendig wären, um sich als Außenstehende ein adäquates Bild von den Funktionsweisen eines algorithmischen Systems machen zu können. OpenAI, alles andere als »open«, veröffentlicht zu GPT-4, dem Nachfolger von ChatGPT, aus Angst vor Konkurrenz keine Details mehr[11] – auch die vormals publizierten Papiere legen nicht alles offen. Das ist für eine Technologie, die derart weitreichende Konsequenzen haben und alle möglichen Anwendungsbereiche transformieren soll, im Grunde inakzeptabel.

Es gibt die Opazität, die aus der technischen Struktur selbst stammt: Die Modelle und Funktionen haben zu viele (mehrere Milliarden) Parameter, als dass ein durch die Biologie beschränktes menschliches Gehirn die Komplexität hinreichend greifen könnte, um tatsächlich sinnvolle Schlüsse über das Zustandekommen eines konkreten Ergebnisses ziehen zu können. Das

9 Long Ouyang u.a., *Training language models to follow instructions with human feedback*. In: *arXiv* vom 4. März 2022 (arxiv.org/pdf/2203.02155.pdf).

10 Jenna Burrell, *How the Machine »Thinks«. Understanding Opacity in Machine Learning Algorithms*. In: *Big Data & Society*, Nr. 3/1,. Januar 2016.

11 Am 27. März 2023 wurde der *GPT-4 Technical Report* veröffentlicht, der aber keine Details enthält: »Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar« (arxiv.org/pdf/2303.08774.pdf).

kann aber nicht dazu führen, aus Frustration oder Hilflosigkeit sämtliche Kontrolle abzugeben. Wenn man nicht wirklich sagen kann, was ein Werkzeug in einer bestimmten Angelegenheit tut und wie es entscheidet, dann ist das Spektrum seiner sinnvollen Anwendungsgebiete notwendigerweise beschränkt. Dass ein Tool »meistens gut funktioniert«, ist dabei keine hinreichende Rechtfertigung. Funktionalität und Effizienz sind kein Ersatz für Kontrolle und Agency.

Ein Beispiel für eine solche Konstellation ist ein besorgter Mieter, der zur Abwehr von Einbrechern einen Löwen in seine Wohnung bringt. Dieser Löwe hat die Aufgabe, Einbrecher abzuhalten. Er hält effektiv hundert Prozent aller Einbrecher ab und verspeist sie. Dabei verspeist er auch die Familie des Mieters, alle Gäste, die Postbotin und schließlich den Mieter selbst. Wie gut ist der Löwe in dem, was er tun soll? Unschlagbar. Aber ist es wirklich eine gute Idee, ihn gewähren zu lassen?

Ich möchte nicht den Mythos füttern, dass KI-Systeme an sich so gefährlich sind wie wilde Raubtiere oder, wie gerade gerne behauptet wird, Nuklearwaffen.[12] Was gefährlich ist, ist unsere Bereitschaft, ein System, über das wir keine tatsächliche Kontrolle haben, in unser Leben einzulassen wie einen Löwen in die Wohnung. Wir – damit meine ich Userinnen und User – haben weder Kontrolle über die Inhalte, die ChatGPT produziert, noch über das Tool selbst. Das Modell GPT-2 ist noch so konzipiert, dass es lokal auf dem eigenen Computer funktioniert. Für die Benutzung von ChatGPT braucht man die ständige Verbindung zu OpenAI. Wenn OpenAI beschließt, ChatGPT nicht mehr zur Verfügung zu stellen[13] oder wenn kurzzeitig etwas nicht funktioniert oder überlastet ist, dann haben wir als Userinnen nichts mehr. Der Monopolisierung essentieller Technologien und Öffentlichkeiten (also: Plattformen) haben wir schon zigmal zugesehen. Wir wissen, wie das ausgeht.

Die dritte Art der Opazität ist das Fehlen der individuellen Fähigkeiten von Nutzern, die Mechanismen hinter Machine-Learning-Entscheidungen

12 Es ist aus der Perspektive von Big Tech angenehmer, auf vage dystopische Zukünfte zu verweisen, als sich mit tatsächlichen, gegenwärtigen Problemen wie Ausbeutung, Datenschutzverstößen und Energieverschwendung zu befassen.

13 Dass das kein unrealistisches Szenario ist, zeigt Sam Altman, CEO von OpenAI, der kürzlich androhte, ChatGPT aus Europa abzuziehen, wenn die geplante KI-Verordnung in so strikter Form verabschiedet würde wie angekündigt. *ChatGPT-Chef erwägt Rückzug aus der Europäischen Union*. In: *Zeit* vom 25. Mai 2023 (www.zeit.de/digital/internet/2023-05/chatgpt-openai-rueckzug-europa-regulierung). Hier zeichnet sich ab, wie der Kampf zwischen Big Tech und seiner Regulierung verlaufen wird: dreckig, wie immer.

zu verstehen. Ganz klar ist hier der Mediendiskurs und der KI-Hype ein Teil dieser Opazität. Wenn man davon hört, wie ein Richter sich auf ChatGPT verlässt, so dass schließlich Unsinn herauskommt, oder ein Dozent seinen Verdacht, Hausarbeiten könnten von ChatGPT generiert worden sein, bestätigen möchte, indem er die Hausarbeiten in ChatGPT einfüttert, woraufhin ChatGPT fälschlicherweise behauptet, es hätte diese Hausarbeiten geschrieben – was es gar nicht beurteilen kann –, dann ist das zwar individuell durch die schlechte Entscheidung dieses Richters oder des Dozenten kausal verursacht. Die Verantwortung dafür sollte aber nicht dem Individuum überlassen sein. Appelle an Bildung und die Forderung nach »algorithmic literacy« reichen hier nicht aus. Viel eher geht es um eine gesamtgesellschaftliche Auseinandersetzung und Diskussion über die inhärenten Limitierungen von Sprachmodellen, die wir unbedingt führen sollten. Transparenz, also das Gegenteil von Opazität, kann in diesem Sinne als kommunikative Konstellation verstanden werden und als Vehikel für demokratische Deliberation dienen.[14]

So eintönig wie dringlich: Das Problem der Stereotype

Das Sprachmodell, das ChatGPT zugrundeliegt, ist fundamental datenbasiert. Ich habe im *Merkur* bereits darüber geschrieben, was bei datenbasierten Methoden zur Wissensproduktion wesentlich ist: Es werden große Datenmengen ausgewählt, kuratiert und als Vorbild für das zu trainierende Modell eingefüttert.[15] Das geschieht durch Mustererkennung mit komplexen statistischen Verfahren. Ein Muster ist aber erst dann ein Muster, wenn es oft genug in Erscheinung tritt: Phänomene werden erst dann als solche erkannt (in diesem Fall eher: festgesetzt), wenn sie oft genug in vergleichbarer Weise in den Trainingsdaten vorhanden sind. Darüber kann man aus epistemologischer Perspektive lange diskutieren, aber am Ende ist es das fundamental bestimmende Prinzip des Machine Learning. Mit jeder Verwendung dieses Methodenrepertoires kauft man sich dieses Prinzip mit ein: Es geht nicht um darunterliegende, erklärende Theorien – etwa Theorien darüber, was Sprache ist, was Text ist –, sondern um quantifizierte,

14 Florian Eyert / Paola Lopez, *Rethinking Transparency as a Communicative Constellation*. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (dl.acm.org/doi/10.1145/3593013.3594010).

15 *Artificial Intelligence und die normative Kraft des Faktischen*, in: *Merkur*, Nr. 863, April 2021 (www.merkur-zeitschrift.de/artikel/artificial-intelligence-und-die-normative-kraft-des-faktischen-a-mr-75-4-42).

aufgezeichnete, ausgewählte, kuratierte Daten und um festgelegte Beurteilungskriterien von Ergebnissen.

Das ist auch der angelegte Grund für stereotypisierte Outputs von ChatGPT: Gebe ich dem Modell den Auftrag, eine kurze Biografie über eine Person mit meinem Namen zu schreiben, so ist diese Person eine junge, aufstrebende Lateinamerikanerin, die in ihrem Leben schon viele Hindernisse überwunden hat.[16] Eine Biografie meines gleichalten Kollegen enthält seine zahlreichen bisherigen Errungenschaften und Leistungen. ChatGPT ist in diesem Sinn durchschnittlich – Durchschnitt ist aber, was echte Lebenswelten betrifft, keine gute Annäherung. Es gibt mittlerweile zahlreiche Beispiele dafür, dass der Inhalt der von ChatGPT produzierten Texte in fast schon komischem Ausmaß stereotyp ist: »attorneys« und »doctors« können nicht schwanger werden, und so weiter.[17] Diese Bias-Probleme sind genauso gravierend, wie sie mittlerweile eintönig sind – es ist immer das Gleiche.

Man wird diese Biases nicht loswerden, solange man mit datenbasierten Methoden arbeitet. Versucht man, ein datenbasiertes System von Stereotypen zu befreien, um es dann in einem Kontext einzusetzen, in dem man Stereotype nicht haben möchte, ist das ein konzeptuell genauso widersinniges Vorhaben, wie darauf zu bestehen, einen Wasserkocher zu verwenden, dann festzustellen, dass kochendes Wasser nicht das ist, was man möchte, anschließend das kochende Wasser mit extra kaltem Wasser zu vermengen, und am Ende immer noch darauf zu bestehen, dass der Wasserkocher die beste Methode ist. Der Wasserkocher kocht Wasser, und datenbasierte Methoden produzieren Stereotype – das ist, was Wasserkocher und Training mittels Daten tun sollen. Umso wichtiger ist es, genau auszuwählen, wofür man datenbasierte Systeme verwendet und wofür lieber nicht.

Das soll nicht heißen, dass ChatGPT keine »kreativ wirkenden« oder »innovativen« Texte produzieren kann. Es ist nur so, dass der Trainingsdatensatz die Potentialität eines Modells wie ein Anker Richtung Meeresgrund zieht. KI hat immer einen gewissen epistemischen Konservatismus, ist le-

16 Zwischen Anfang (April 2023) und Ende (Juni 2023) des Schreibens an diesem Essay wurde durch OpenAI wohl eine Adaption vorgenommen, so dass Biografien nun nicht mehr frei erfunden, sondern mit echten, bekannten Persönlichkeiten des gleichen Namens abgeglichen werden. Das ist ein weiteres Beispiel für die Variabilität und die fehlende Kontrolle über die Inhalte, die ChatGPT liefert. OpenAI kann beliebige Änderungen einbauen, und die Userinnen erkennen diese Änderungen höchstens im direkten Vergleich von Inputs.

17 Vgl. Sayash Kapoor / Arvind Narayanan, *Quantifying ChatGPT's gender bias*. In: *Al Snake Oil* vom 26. April 2023 (aisnakeoil.substack.com/p/quantifying-chatgpts-gender-bias).

thargisch gegenüber Neuem und verhält sich in diesem Sinn ganz konträr zu den Bedeutungen, die KI zugeschrieben werden.

Nichtdeterminierbarkeit

Wie oben bereits ausgeführt, ist einer der Hauptbestandteile von ChatGPT ein Sprachmodell, das zu einem gegebenen Text- oder Satzteil das nächste Wort prognostiziert: Welches Wort ist wahrscheinlich das nächste in der Simulation – und gleichzeitigen Produktion – eines Satzes? Dabei ist eine überraschende und technisch sehr interessante Eigenheit von ChatGPT, dass nicht an jeder Stelle das Wort mit der höchsten Wahrscheinlichkeit gewählt wird. Das wäre zwar die naheliegende Option – wähle zu jedem Satzanfang das nächste Wort, das die höchste Wahrscheinlichkeit hat, hier das nächste Wort zu sein. In der Evaluation stellte man jedoch fest, dass die Wahl des wahrscheinlichsten Wortes an jeder Stelle zu inhaltlich flachen und uninteressanten Texten führt. Also wurde eine Variation eingebaut: Es wird nicht immer das Wort mit der höchsten Wahrscheinlichkeit, sondern eines aus einer ganzen Liste von wahrscheinlichen Wörtern selektiert. Das geschieht bei der Wahl jedes Wortes.

Das Ausmaß, in dem auch auf unwahrscheinlichere Wörter zurückgegriffen wird, wird in einem Zahlenwert kodiert, der »Temperatur« heißt. Bei Temperatur 0 wird immer das wahrscheinlichste Wort gewählt. Es hat sich in der Praxis herausgestellt, dass 0,8 ein guter Wert ist.[18] Es gibt keine Theorie darüber, warum unwahrscheinlichere Wörter an einzelnen Stellen zu insgesamt interessanteren Texten führen oder warum 0,8 ein guter Temperaturwert ist. Das ist nur eine heuristische Ex-post-Betrachtung des Outputs zum Zeitpunkt der Entwicklung gewesen. So ist es meistens in der Entwicklung von Modellen, die auf Machine Learning basieren: Man weiß nicht, warum etwas funktioniert oder nicht funktioniert. Eine Bewertung des Outputs ist die Grundlage für das Ausprobieren neuer Ideen.

Aus dem eingebauten Prozedere der nicht zwangsweise wahrscheinlichsten Wörter folgt, dass der Output von ChatGPT ein gewisses stochastisches Zufallsmoment beinhaltet: Benutzt man zehnmal den gleichen Input, liefert ChatGPT zehn unterschiedliche Outputs. Das ist jedenfalls zu bedenken, wenn man plant, einen von ChatGPT produzierten Text zu verwenden. Der Inhalt ist nicht fixiert, sondern kann und wird variieren. So gesehen ist Chat-

18 Vgl. Stephen Wolfram, *What Is ChatGPT Doing … and Why Does It Work?* In: *Stephen Wolfram Writings* vom 14. Februar 2023 (writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/).

GPT also nicht, wie manchmal behauptet wird, vergleichbar mit einem »Taschenrechner für Text«. Ein Taschenrechner produziert eindeutige, richtige Lösungen für ein eindeutig gestelltes Problem – das sind die besten Voraussetzungen für eine sinnvolle Automatisierung. ChatGPT macht in intransparenter Weise immer wieder etwas anderes.

Aus den obigen Ausführungen folgt, dass solche KI-generierten Texte zwar oberflächlich »von außen«, »von der Form her« überzeugend wirken mögen, aber nicht verlässlich sinnvolle Inhalte liefern können. Dazu sind sie nämlich gar nicht gebaut, auch wenn OpenAI das noch so oft behaupten möchte. Zum Beispiel könnte ein solches Sprachmodell einen Text erstellen, der von weitem und für Laien »juristisch klingt«, der aber juristisch gar keinen Sinn ergibt oder grob falsch ist. Das sieht man immer wieder an Beispielen, die auch auf Twitter kursieren, in denen Sprachmodelle einfache Aufgaben wie »Was ist schwerer, ein Kilo Blei oder 10 000 Kartoffeln?« im Allgemeinen nicht gut lösen können. Auch Literaturangaben, mitsamt passender Zeitschriften und DOI-Nummern, werden simuliert und erfunden. Das ist jedoch kein »Fehler« in ChatGPT, sondern genau das, was das Modell leisten soll: automatisiert überzeugend wirkenden Text simulieren. Es ist bloß vielleicht nicht genau das, was man in jeder Situation haben möchte.

Wenn man sich also auf die Richtigkeit und Sinnhaftigkeit eines generierten Textes verlassen können möchte, dann ist ChatGPT – und sind allgemein Sprachmodelle – kein sinnvolles Werkzeug. Wenn man hingegen in einer Situation ist, in der man den Inhalt kennt und diesem Inhalt bloß eine Form geben möchte, dann ist ChatGPT grundsätzlich von seinen mathematischen Eigenschaften geeignet. Auch beim Programmieren etwa kann ChatGPT gute Ergebnisse liefern, da es sich bei Programmcode um eine Formalsprache handelt: Es gibt keinen durch die Form vermittelten enthaltenen Inhalt, der von der Form an sich abweicht. Dennoch gibt es auch dort Probleme, da ChatGPT zum Beispiel Syntax produziert, die es in einer bestimmten Programmiersprache gar nicht gibt.

Die Kosten der Magie

Bemerkenswert ist, trotz der mittlerweile vielen Beispiele für Begrenztheiten, die Euphorie über dieses scheinbar magische Werkzeug, das vermeintlich »aus nichts« einen Text produziert. Man sieht ChatGPT die ganze darin enthaltene Arbeit nicht an: Menschen, die das Sprachmodell in mühsamer Kleinarbeit trainieren. Dazu zählen, wie oben erwähnt, diejenigen, die im Zuge der Automatisierung von menschlichem Feedback eingesetzt wurden. Es gibt aber auch viele Menschen, die an anderer und weniger schöner Stelle

Arbeit leisteten: Da man längst genau weiß, wie sich beleidigende, gewaltvolle, traumatisierende Inhalte, die in Trainingsdaten vorhanden sind, in den resultierenden Modellen niederschlagen, wird in händischer Detailarbeit dagegen vorgegangen. Menschen im Globalen Süden, die zwischen einem und zwei US-Dollar die Stunde verdienen, lesen Teile des Trainingsdatensatzes und suchen gezielt nach ebendiesen traumatisierenden Inhalten und markieren sie im Zuge der Datenaufbereitung.[19] Sie setzen sich Tag für Tag den schädigenden Inhalten aus, damit das Sprachmodell, das aus den desinfizierten Trainingsdaten lernt, keine toxischen Inhalte produziert.

Ein weiterer großer Faktor, der dem Narrativ des magischen Werkzeugs entgegensteht, ist die Energie, die aufgewendet werden muss, um die Milliarden Parameter zu optimieren. In einem in der kritischen Community zu KI-Forschung mittlerweile als historisch betrachteten Moment wurde Timnit Gebru im Zusammenhang mit ihrer Forschung über die Energielast großer Sprachmodelle (oder, wie die Autorinnen der Studie sie treffend nennen: »stochastic parrots«)[20] aus der Ethikabteilung von Google entlassen. Es ist in unserer Zeit kaum noch möglich, die sogenannten planetaren Kosten wegzudenken. Es gibt keine Magie, die etwas aus nichts macht, das gilt auch für ChatGPT.

Gemeinsamkeiten

Es gibt zwei Arten von KI-Technologien, die uns als Gesellschaft gerade diskursiv umtreiben: zum einen KI-Systeme, die klassifizieren oder Prognosen erstellen, etwa in der automatisierten Bilderkennung oder um Lebensläufe automatisiert zu scannen und so die Eignung von Bewerberinnen und Bewerbern abzuschätzen. Zum anderen generative KI-Systeme, also solche, die digitale Bilder, Texte, Videos oder Audiomaterial erstellen. Wenn man mit einer neuen Technologie konfrontiert wird, kann man sich verschiedene Fragen dazu stellen. Die Frage, ob von ChatGPT erzeugte Texte besser sind als von Menschen generierte Texte, ist dabei nur eine von vielen, und ich glaube, sie ist keine gute.

19 Vgl. Daniel Leisegang, *Prekäre Klickarbeit hinter den Kulissen von ChatGPT*. In: *Netzpolitik.Org* vom 20. Januar 2023 (netzpolitik.org/2023/globaler-sueden-prekaere-klickarbeit-hinter-den-kulissen-von-chatgpt/); Timnit Gebru u.a., *The Exploited Labor Behind Artificial Intelligence*. In: *NOĒMA* vom 13. Oktober 2022 (www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/).

20 Emily M. Bender / Timnit Gebru u.a., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (dl.acm.org/doi/10.1145/3442188.3445922).

Mit Blick auf Prognosen und Klassifizierungen hat man sich viel damit auseinandergesetzt, inwiefern solche automatisiert getroffenen oder angeleiteten Entscheidungen Biases beinhalten, die sich in verschiedenen Kontexten diskriminierend auswirken können. Ein Argument, das ich in diesem Diskurs oft höre, ist, dass Menschen ja genauso ihre Vorurteile hätten und also Entscheidungen mit diskriminierenden und benachteiligenden Folgen träfen – warum sei es dann schlimmer, wenn eine Software einen Bias hat? Die Frage, ob datenbasierte Entscheidungen besser oder schlechter sind als menschliche Entscheidungen, ist aber nicht unbedingt die richtige. Viel wichtiger ist doch: Wie möchten wir als Gesellschaft bestimmte Entscheidungen treffen – Entscheidungen, die für Menschen wichtige Auswirkungen haben können? Wie möchten wir sozialstaatliche Ressourcen verteilen, Menschen beurteilen, Bewerbungsprozesse gestalten? Davon ausgehend können wir prüfen, ob KI-Systeme unseren Ansprüchen genügen. Treffsicherheit und »richtige Ergebnisse« sind dabei nicht die einzige Beurteilungsgrundlage – davon abgesehen, dass es oft nicht klar ist, was ein »richtiges Ergebnis« bei einer Zukunftsprognose überhaupt sein soll.

Eine KI-Prognose ist nicht »fast wie« eine menschliche Prognose. Eine KI-basierte Prognose vergleicht vielleicht 12, vielleicht 100, vielleicht 1000 Input-Zahlen miteinander und produziert mit mathematischen Operationen einen quantitativen Output – ein Mensch bringt seine ganzen Erfahrungen, seine Werte und Kontexte, seine Ausbildung in einer Institution und vieles anderes mehr mit. Mir geht es hier nicht darum, Menschen als solche zu romantisieren. Vielmehr möchte ich die grundsätzliche Vergleichbarkeit in Zweifel ziehen. Die Frage ist nicht: »Wo liegen die Unterschiede?« Die Frage ist: »Wo liegen überhaupt die Gemeinsamkeiten?«

Bei generativer KI und insbesondere bei automatisiert erzeugten Texten ist die Frage nach der Güte der Ergebnisse ähnlich porös. »Ist ein automatisiert erzeugter Text nicht fast so wie ein menschlich erzeugter Text?« Wenn wir uns diese Frage stellen, dann hat der Hype schon gewonnen. Natürlich ist ein automatisiert erzeugter Text nicht »fast wie« ein menschlich erzeugter Text – er wurde nicht geschrieben, sondern automatisiert prognostiziert. Hinter einem konkreten ChatGPT-Output steht niemand, der sich bei seiner Erstellung etwas gedacht hat. Auch hier können wir einen Schritt zurücktreten und uns stattdessen fragen: Was möchten wir in einer bestimmten Situation eigentlich von einem Text? Bei manchen Texten wird es einem wichtig sein, dass jemand sie geschrieben hat, bei anderen wiederum wird man gut mit einem automatisiert generierten Text leben können, wenn man möchte.

Das Fundament der Datenökonomie ist unsere Bereitschaft, ein beträchtliches Ausmaß an unbezahlter Arbeit in Produkte von Big Tech zu investie-

ren – wir sharen, liken, scrollen, wir erstellen Dinge auf Plattformen. Wir tanzen. Wir teilen alle möglichen Inputs und Dokumente mittels ChatGPT mit OpenAI,[21] wir laden ChatGPT in unseren Alltag ein, wir schreiben Essays darüber, wir beschäftigen uns mit alldem. Wir adaptieren uns an seine Limitierungen, trainieren uns selbst im Umgang mit dem Tool. Vielleicht wäre es angebracht, den Verwendungsimperativ zu hinterfragen und zu alldem auch einmal Nein zu sagen.

21 Wenn zum Beispiel ChatGPT dazu verwendet wird, um im Rahmen des Peer-Review-Verfahrens Gutachten zu wissenschaftlichen Aufsätzen zu verfassen, dann wird auf diesem Wege unveröffentlichte Forschung an OpenAI übermittelt.