Wie lange braucht es uns noch?

Maschinen können seit neustem Gespräche führen, Bilder erschaffen – und Artikel schreiben. Die Folgen für die Menschheit sind so dramatisch wie unabsehbar.

Von Reto U. Schneider, NZZ Folio, 05.09.2022

Ein Computer wird nie in der Lage sein, einen Text zu schreiben, weil er die Nuancen der menschlichen Sprache nicht verstehen kann. Computer sind gut darin, präzise Anweisungen zu befolgen, aber sie können die Feinheiten und die Komplexität der menschlichen Kommunikation nicht erfassen. Ein Computer wäre zum Beispiel nicht in der Lage, zwischen den Bedeutungen von «ich liebe dich» und «ich hasse dich» zu unterscheiden.

Halten Sie den vorangegangenen Abschnitt für einen gelungenen Einstieg? Falls nicht, trifft mich keine Schuld. Ich habe nämlich kein Wort davon geschrieben. Ich habe lediglich die Website beta.open−ai.com/playground aufgerufen und dort die Aufforderung eingetippt: «Schreibe einen Text darüber, warum ein Computer niemals in der Lage sein wird, einen Text zu schreiben.» Das Resultat haben Sie eben gelesen. Ersonnen hat die drei Sätze die im Mai 2020 lancierte künstliche Intelligenz GPT-3, die dritte Generation des Generative Pretrained Transformers des KI-Unternehmens OpenAI mit Sitz in San Francisco. GPT-3 hat sie nicht im Internet abgeschrieben oder leicht verändert übernommen, nein, GPT-3 hat sie exklusiv für mich geschmiedet. Und auch auf die nächste Frage textet sie eine individuelle Antwort:

«Was ist der Unterschied zwischen ‹ ich liebe dich › und ‹ ich hasse dich ›?»

Es gibt einen grossen Unterschied zwischen «ich liebe dich» und «ich hasse dich». Wenn man «ich liebe dich» sagt, bedeutet das, dass man sich um die Person sorgt und möchte, dass sie glücklich ist. Wenn man «ich hasse dich» sagt, bedeutet das, dass man die Person nicht mag und will, dass sie unglücklich ist.

Im Schatten von Corona und dem Ukrainekrieg hat die künstliche Intelligenz eine Fähigkeit erlangt, die wir bisher für typisch menschlich hielten: den Umgang mit Sprache.

«Unsere Sprache macht uns am meisten aus», sagt der Informatiker Richard Socher, dessen Forschung vor zehn Jahren entscheidend dazu beigetragen hat, dass wir uns heute mit Computern unterhalten können. Socher ist der hochbegabte Sohn eines Wissenschafters und einer Ingenieurin aus der früheren DDR. Nach seinem Studium in Deutschland ging er in die USA und lehnte als 31jähriger eine Professur an der Eliteuniversität Princeton ab, um ein Start-up-Unternehmen zu gründen – und damit reich zu werden. 2020 rief Socher you.com ins Leben, eine Suchmaschine, die eine bessere Kontrolle der Privatsphäre und der Wahl der Quellen verspricht.

Etwas versteckt findet sich auf you.com auch Youwrite, ein «Schreibassistent, der mit künstlicher Intelligenz ausgestattet ist», wie es in der Eigenwerbung heisst. Youwrite benutzt im Hintergrund GPT-3. Der digitale Assistent soll den Schreibstil verbessern oder gegen Schreibhemmung helfen. Schüler haben längst entdeckt, dass er auch gegen Hemmungen hilft, die Hausarbeiten zu erledigen. Youwrite schreibt in wenigen Sekunden kleine Aufsätze zu selbstgewählten Themen, die sich einzig dadurch verdächtig machen, dass sie oft besser sind, als es von den Schülern erwartet werden kann. Im Gegensatz zum Abschreiben haben diese Texte aus Schülersicht den Vorteil, dass keine Plagiatssuche sie je entdecken wird – ganz einfach, weil sie keine Plagiate sind.

Die Möglichkeit, an der Schule oder Uni unentdeckt zu betrügen, ist nur eine Folge, die das Sprachverständnis der Maschinen nach sich zieht. Andere betreffen

Werbetexter und Übersetzerinnen, Schriftstellerinnen und Dichter − und alle übrigen, die davon leben, aus Worten Sätze zu bauen, auch Journalisten wie mich.

Derzeit ist es noch schwer vorherzusagen, inwieweit die künstliche Intelligenz die journalistische Arbeit beeinflussen wird. Möglicherweise werden einige Aufgaben, die bisher von Journalisten erledigt wurden, künftig von künstlichen Intelligenzen übernommen. Dies könnte dazu führen, dass einige Journalisten ihren Job verlieren.

Ups! Wie alle kursiven Passagen in diesem Artikel hat den vorangegangenen Text GTP-3 geschrieben. Ich habe sie gefragt, welche Folgen das Sprachvermögen der künstlichen Intelligenz für den Journalismus habe. Wenn sie mir schon den Job stiehlt, kann sie mir wenigstens noch bei diesem Artikel helfen.

Manche Forscher spielen die Erfolge herunter. Sie glauben, wenn künstliche Intelligenz «prädiktive Optimierung mittels geschichteter Regressionen» hiesse, würde sich kaum jemand dafür interessieren. Andere sind überzeugt, dass die zwanziger Jahre des 21. Jahrhunderts dereinst nicht wegen einer Pandemie und einem Krieg in die Geschichte eingehen werden, sondern als die Zeit, in der die Maschinen die Sprache lernten. Natürlich gibt es kommerziell wichtigere Anwendungen der künstlichen Intelligenz, etwa die Bilderkennung oder Programme, die die Faltung eines Proteins berechnen. Aber keine ist so eng mit unserem Leben verflochten wie die Sprache.

«Wir sind im Elektrizitätsstatus», sagt Socher. Als der Mensch die Elektrizität beherrschen lernte, veränderte er damit die Welt. Die Gaslampen wurden ersetzt, Maschinen neu mit Strom betrieben. «Ähnlich ist es jetzt mit der künstlichen Intelligenz.» Anstelle von Strom werden die Menschen aus Riesenrechnern künstliche Intelligenz beziehen, die «repetitive intelligente Aufgaben» übernehme, sagt Socher. Man braucht kein Experte zu sein, um vorherzusagen, dass es nicht bei repetitiven Aufgaben bleiben wird. Einst ersetzte die Dampfmaschine die Muskelkraft. Jetzt ereilt die Geisteskraft ein ähnliches Schicksal.

Schon heute kann sich eine künstliche Intelligenz wie GPT-3 nicht nur geistreich mit Menschen unterhalten oder strukturierte Argumente liefern. Andere Programme sind in der Lage, nach Textbeschreibungen hochauflösende Bilder zu erzeugen, ganz egal wie absurd die Vorgabe klingt. Der Auftrag «Bilder der Überwachungskamera von Cäsars Ermordung» bringt grobkörnige Schwarzweissfotos zurück, auf denen schemenhaft Männer in Tuniken zu sehen sind. Man kann einen fotorealistischen Eisbären in der Wüste verlangen oder eine griechische Statue, die über eine Katze stolpert. Stilistisch gibt es keine Grenzen: Kinderzeichnung, Picasso, Polaroidbild. Auch wenn das Wort im Zusammenhang mit Computern schwer zu definieren ist: die Maschinen scheinen zu verstehen.

Die Konsequenzen dieses Durchbruchs sind so dramatisch wie unvorhersehbar, denn die Benutzung der Sprache unterscheidet sich von jeder anderen Fähigkeit des Menschen. «Natürlich haben Tiere auch Sprache», sagt Socher, aber an Komplexität sei sie jener der Menschen unterlegen. Wir können über Äpfel und Ängste sprechen, über vergangene Schlachten und die zukünftigen Kindergeburtstage, über die Waschmaschine und Jupiter. Wir können ausdrücken, was es gibt, noch nicht gibt oder nie geben wird. «Sprache ist das Blut der Seele, in die Gedanken fliessen und aus der sie entstehen», hat der amerikanische Gelehrte Oliver Wendell Holmes gesagt. Deshalb ist die Sprache für die künstliche Intelligenz so wichtig. Sie kommt dem, was wir Denken nennen, am nächsten. Oder wie es Socher sagt: «Sprache ist die interessanteste Manifestation menschlicher Intelligenz.»

Davon war auch der britische Mathematiker Alan Turing überzeugt. *In seiner legendären Abhandlung «Computing Machinery and Intelligence» von 1950 stellte er die Frage: «Kann eine Maschine denken?» Dazu formulierte er den nach ihm benannten «Turing-Test». Danach wäre eine Maschine dann intelligent, wenn sie einen Menschen davon überzeugen könnte, dass er mit einem anderen Menschen spricht und nicht mit einer Maschine. Das klingt vielleicht einfach, aber es ist es nicht.*

Denn was macht einen Menschen aus? Und was ist mit dem Konzept des Bewusstseins? Wie kommt eine Maschine darauf, was ein Mensch denkt und fühlt?

Um den kursiven Text oben zu erhalten, brauchte ich GTP-3 noch nicht einmal eine Frage zu stellen. Ich habe nur den vorangegangenen Absatz ins Textfeld kopiert. Die künstliche Intelligenz schreibt dann einfach weiter und nimmt dabei die vorangegangenen Gedanken auf.

Jahrzehntelang konnten die Maschinen die Menschen nicht täuschen. Doch heute, *dank GPT-3, ist das anders. Die KI ist in der Lage, menschliche Sprache so gut zu verstehen und zu interpretieren, dass sie in der Lage ist, einen komplett neuen Satz zu generieren, der sich nahtlos an den vorherigen anschliesst.*

Selbst hartgesottene KI-Forscher, die genau wissen, wie die Maschinen aufgebaut sind, müssen sich immer wieder sagen: *«Das ist unglaublich.» GPT-3 ist ein Meilenstein in der Geschichte der künstlichen Intelligenz und wird die Art und Weise, wie wir künftig mit Maschinen interagieren, grundlegend verändern.* So hört sich also Eigenwerbung einer künstlichen Intelligenz an.

Die Geschichte der künstlichen Intelligenz ist auch die Geschichte spektakulärer Misserfolge. Als das Forschungsgebiet 1956 an einer Fachkonferenz geboren wurde, herrschte Optimismus. Bald konnten erste Programme Dame spielen oder lösten mathematische Textaufgaben.

Aber es zeigte sich, dass die Wissenschafter sich in einem entscheidenden Punkt getäuscht hatten. Sie glaubten, was Menschen leichtfalle, würde auch Computern leichtfallen. Dabei war es gerade umgekehrt. Die Maschinen entwickelten sich zwar zu passablen, später sogar zu meisterhaften Schachspielern, aber *sie konnten nicht einmal ein einfaches Gespräch führen.* Bravo GPT-3, genau das wollte ich sagen.

Die Phasen der Stagnation zogen sich so lange hin, dass sie in der Forschung KI-Winter genannt wurden. Fortschritte gab es erst, als man sich von der Idee

verabschiedete, dass der Mensch dem Rechner alle Regeln vorkauen und eigenhändig in ein Programm schreiben muss. Neu sollten die Computer selber lernen − wie Kinder. Eigentlich wusste man schon lange, wie das ging. Bereits 1958 hatten zwei Forscher ein sogenanntes Perzeptron gebaut, das anhand von Beispielen lernte, Muster zu erkennen.

Auf der Grundlage dieser Idee wurden später sogenannte neuronale Netzwerke entwickelt, deren Innenleben im ersten Moment recht einfach erscheint. Sie bestehen aus in Schichten angeordneten Schaltstellen − sogenannten künstlichen Neuronen − und haben eine gewisse Ähnlichkeit mit dem Gehirn. Jedes dieser Neuronen ist über viele Verbindungen mit den Neuronen einer Schicht darunter und einer Schicht darüber in Kontakt. Die Tätigkeit eines einzelnen Neurons besteht in der einfältigen Aufgabe, alle Signale der eingehenden Verbindungen der Schicht darunter zusammenzuzählen, und wenn die Summe einen bestimmten Wert überschreitet, an die Neuronen in der Schicht darüber weiterzugeben. Dort geschieht bei jedem Neuron wieder dasselbe: zusammenzählen und abhängig von der Summe ruhig bleiben oder selber feuern. Das ist alles.

Bahnbrechend an dieser Idee ist, dass ein solches Netzwerk anhand von Beispielen lernen kann, indem es die Durchlässigkeit jeder einzelnen Verbindung zwischen zwei Neuronen verändert. Die Forscher sprechen vom Gewicht einer Verbindung. Man füttert die KI zum Beispiel mit Textteilen und lässt jeweils das letzte Wort ___. Dann werden die Gewichte der Verbindungen so lange verändert, bis das Netzwerk das richtige Wort − in diesem Fall «weg» − ausgibt. Nach Millionen solcher Beispiele kann die Maschine irgendwann ohne Hilfe die wahrscheinlichsten nächsten Worte erraten. Der Vorgang ist der Technik ähnlich, die im Google-Suchfeld «Monroe» vorschlägt, wenn man «Marilyn» eingetippt hat.

Auf diese Weise hat auch die künstliche Intelligenz GPT-3 trainiert, bis sie der mächtigste Autovervollständiger war, den die Welt je gesehen hat. Im Jargon heisst

ein solches Programm Large Language Model − grosses Sprachmodell. Als sie im Frühling 2020 vorgestellt wurde, gerieten auch Fachleute ins Staunen. GPT-3 beantwortete Fragen, schrieb Gedichte und vollendete Artikel. Die Texte hatten nichts Mechanisches, es gab keine Wiederholungen, kaum grammatikalische Fehler.

Dass eine solche Sprachgewalt allein durch das Erraten des jeweils nächsten Wortes zustande kommen sollte, war schwer zu glauben. Doch so war es. Obwohl man das Gegenteil schwören könnte, hat GPT-3 keine Ahnung, wer Alan Turing ist, und weiss nicht, was Liebe bedeutet. Das Geheimnis von GPT-3 ist, dass es keines gibt. Es brauchte kein neues Verfahren, keine revolutionäre Entdeckung, es war die schiere Grösse, die GPT-3 ihre einzigartigen Fähigkeiten verlieh.

Die künstliche Intelligenz GPT-3 hat 175 Milliarden Verbindungen zwischen ihren Neuronen, die bei jedem Trainingsbeispiel so lange justiert wurden, bis das fehlende Wort am Ausgang erschien. Wenn das Einstellen einer Verbindung eine Sekunde dauern würde, hätte ein einziges Trainingsbeispiel 5500 Jahre in Anspruch genommen − und GPT-3 hat an 300 Milliarden Beispielen geübt.

Lange Zeit fehlte dazu die Rechenleistung, und es gab nicht genügend Übungsbeispiele. Heute sind die Rechner schnell genug, und das Internet bietet ein unerschöpfliches Reservoir an Trainingsmaterial. GPT-3 wurde an allen sechs Millionen Wikipedia-Artikeln trainiert, was gerade mal drei Prozent der Übungsbeispiele ausmachte. Hinzu kamen Millionen von Büchern und Websites. Falls Sie je etwas im Internet veröffentlicht haben, ist es gut möglich, dass GPT-3 auch daran geübt hat.

«Nur die grossen Firmen haben die Infrastruktur, solche Modelle zu trainieren», sagt der Informatiker Jan Milan Deriu von der Zürcher Hochschule für Angewandte Wissenschaften in Winterthur. Deriu forscht am Zentrum für Künstliche Intelligenz an neuronalen Netzen zur Sprachverarbeitung. Er hat schon als Kind versucht, einen Lego-Roboter zu bauen, der sein Zimmer aufräumt, und die Faszination für Computer

hat seither nicht nachgelassen. Nachdem er an der ETH studiert und an der Universität Zürich doktoriert hatte, nahm er die Stelle in Winterthur an. Wie nervenaufreibend die Aufzucht einer künstlichen Intelligenz sein kann, weiss er aus eigener Erfahrung. «Das Training eines solchen Modells ist die Hölle. Wir nennen es auch Babysitten», sagt er, «es ist wie bei einem Fussballspiel. Man kontrolliert aus der Ferne auf dem Handy ständig das Resultat und denkt immer wieder, ‹ was habe ich da bloss für einen Fehler gemacht ›.»

Bei GPT-3 dürfte das Training mehrere Monate gedauert haben. Allein die Kosten dafür werden auf fünf Millionen Dollar geschätzt. Der Betreiber OpenAI wurde 2015 von Tech-Investoren − darunter Elon Musk von Tesla − als gemeinnützige Organisation gegründet mit dem Ziel, «dafür zu sorgen, dass künstliche allgemeine Intelligenz der gesamten Menschheit zugute kommt». Weil sich dieses Ziel als kostspieliger erwies als angenommen, wurde OpenAI 2019 in ein gewinnorientiertes Unternehmen umgewandelt, bei dem jedoch der maximale Gewinn der Aktionäre auf das Hundertfache ihrer Investition beschränkt ist. Wegen drohender Interessenkonflikte stieg Musk aus. Der Hauptgeldgeber ist Microsoft.

Auch Jan Deriu hat GPT-3 ausprobiert. Und auch er erlebte immer wieder Momente, in denen er dachte: «Das kann jetzt nicht sein.» Nach unserem ersten Gespräch schicke ich ihm 200 meiner Artikel. GPT-3 lässt sich nämlich mit eigenen Übungsbeispielen individuell anpassen. Deriu lädt meine Texte hoch, und zwei Stunden später ist Retobot einsatzbereit: Eine KI-Maschine, die in meinem Stil schreibt.

Ich träume schon davon, dass ich im Büro einfach auf den Knopf drücke und GPT-3 *einen meiner Artikel vor die Linse hält, und dann liest er mir etwas vor, was ich nie geschrieben habe. Was ich schreibe, wird irgendwo in den Tiefen seines neuronalen Netzes vergraben sein. Ich werde einen Geistesblitz haben und sagen: «Das ist gut, das kopiere ich», und dann schreibe ich es ab. So einfach wird das sein.*

Welche Merkmale meiner Art zu schreiben GPT-3 aus den 200 Artikeln aufgesogen hat, ist anhand des Texts oben schwer zu beurteilen, aber ich finde das Resultat ganz gelungen. Er scheint mir auf jeden Fall weniger trocken zu sein als die reine GPT-3.

Deriu kann verstehen, dass gewisse Benutzer glauben, die grossen Sprachmodelle hätten ein Bewusstsein. «Es ist eine derart tolle Illusion.» Im vergangenen Juni behauptete der Google-Mitarbeiter David Lemoine, die künstliche Intelligenz Lambda, an der Google arbeitet, sei lebendig. Er wurde freigestellt und später entlassen.

Lemoine kam nach Gesprächen über Freundschaft und Tod mit Lambda zu seinem Urteil. Der Grossteil der KI-Forscher halten seine Ansicht für Unfug. «Es gibt absolut keinen Grund für uns, Zeit mit der Frage zu verschwenden, ob irgendetwas, das irgendjemand im Jahr 2022 zu bauen weiss, empfindungsfähig sei. Es ist es nicht», schrieb der Psychologe und KI-Entrepreneur Gary Marcus von der New York University. Tatsächlich gibt es im Moment wenig Anhaltspunkte, dass die Maschine mehr ist als «eine Tabellenkalkulation für Wörter», wie es Marcus nennt.

Die Versuchung, den grossen Sprachmodellen eine Seele zuzuschreiben, entstammt einem urmenschlichen Impuls: Was tut wie ein Mensch, behandeln wir wie einen Menschen. Selbst KI-Experten vergessen immer wieder, dass sie es mit Maschinen zu tun haben.

Andererseits stellt sich die ketzerische Frage, wieweit auch der Mensch nur eine Maschine ist. Lässt sich mit immer grösseren neuronalen Netzen und immer mehr Trainingsbeispielen menschenähnliche Intelligenz erreichen? Darüber wird gerade heftig gestritten.

Zwischen der Position, die Sprachmodelle seien bloss «zufallsgetriebene Papageien», die nach statistischen Regeln geordnete Wortsequenzen erzeugen, und der

Ansicht, sie seien lebendig und dürften nicht ausgeschaltet werden, liegt allerdings ein weites Feld.

Wer GPT-3 zum Beispiel Kreativität abspricht, weil die Maschine ja bloss ihre Übungsbeispiele neu verquirlt ausspuckt, muss sich die Frage gefallen lassen: Ist nicht genau das eine Definition von menschlicher Kreativität − Neues zu schöpfen, indem wir frühere Erfahrungen und Eindrücke auf überraschende Weise verbinden? Den Auftrag «schreibe einen Nachruf auf einen Nagel» beendete GPT-3 mit dem Satz: *Er wird überlebt von seiner Cousine, der Schraube.*

Die Frage nach der Kreativität stellt sich noch viel dramatischer bei einer anderen künstlichen Intelligenz. Im April 2022 stellte OpenAI Dall-E 2 vor, ein Programm, das Bilder zu Texten erzeugt. Der Name ist eine Anspielung auf den Maler Salvador Dalí und den Roboter Wall-E aus dem gleichnamigen Pixar-Film. Im Befehlsfeld von Dall-E 2 kann man zum Beispiel eintippen: «Ein Roboter, der ein Buch liest» oder «ein altägyptisches Gemälde, auf dem ein Streit darüber dargestellt ist, wer dran ist, den Müll rauszubringen». Sekunden später liefert die KI eine Auswahl von vier hochauflösenden Bildvarianten, die den Beschreibungen entsprechen. Auf maximal 400 Zeichen kann man jede absurde Situation in jedem entlegenen Stil verlangen: Höhlenkunst, römisches Mosaik, Picasso, eine Fotografie mit einer 200-Millimeter-Objektivbrennweite auf einem abgelaufenen Kodakchrome-Rollfilm. Oder man gibt ein Gedicht ein und staunt darüber, wie die künstliche Intelligenz es illustriert.

Wie bei GPT-3 die Texte sind bei Dall-E 2 die Bilder Originale. Sie werden weder kopiert noch aus festen Elementen vorhandener Bilder zusammengesetzt. Jedes Bild ist eine einzigartige Kreation, die es zuvor nie gegeben hat. Dall-E 2 hat von 650 Millionen Beispielen gelernt, den Inhalt von Bildern mit ihrer Beschreibung zu verbinden. Wie bei GPT-3 geschah das durch das Justieren mehrerer Milliarden Verbindungen des neuronalen Netzwerks.

Das Resultat begeistert − und erschüttert − Grafikerinnen und Künstler. Wenn sie in Spasswettkämpfen gegen die Maschine antreten, ist das wie das Rennen zwischen der Schildkröte und dem Hasen. Um auch nur den Hauch einer Chance zu haben, müssen die Menschen einen demütigenden Startvorteil in Anspruch nehmen: Sie brauchen zwei Tage für einen Entwurf, Dall-E 2 zehn Sekunden.

Der Industriedesigner John Mauriello zeigt in einem Youtube-Video 400 Schuhkonzepte, die er mit einer KI in zwei Stunden kreiert hat. Er habe sein ganzes Berufsleben der Verfeinerung seines Handwerks als Designer gewidmet, sagt der Lehrbeauftragte am California College of the Arts, und diese künstliche Intelligenz habe in wenigen Sekunden mehr realitätsnahe Ideen entwickelt, als er in mehreren Tagen oder Wochen schaffen könnte. «Das erste Mal, als ich diese Werkzeuge benutzte, war ich so überwältigt, dass ich nicht schlafen konnte.» Dass die Branche Umwälzungen erwarten, steht für ihn ausser Frage.

«Die Qualität der Idee ist das, was am Ende zählen wird, nicht die technischen Fähigkeiten, sie umzusetzen», sagt Mauriello. Er ist überzeugt, dass in den nächsten paar Jahren mehr Kunst- und Designkonzepte entstehen werden als im ganzen vergangenen Jahrhundert. «Weil KI-Design so schnell geht, bin ich bereit, auch riskantere Ideen zu verfolgen.» Die Grundlage der Maschinen sei «ein grösserer Katalog aller kulturgeschichtlichen und künstlerischen Bewegungen, als jemals ein Mensch in seinem Kopf haben wird». Natürlich stänkern einige Pedanten, dass die Maschinen dieses oder jenes nicht könnten. Mauriello findet, das sei, wie wenn man die Schlüssel für das modernste Raumschiff der Welt bekomme und sich dann beschwere, dass es darin nicht genug Getränkehalter gebe.

Die Diskussion darum, ob es Kunst sei, Worte einzutippen, die eine KI zu einem Bild macht, hat bereits begonnen. Während manche Dall-E 2 für eine nette Spielerei halten, glauben andere, die künstliche Intelligenz werde den Kunstbetrieb stärker durchrütteln als die Fotografie seinerzeit. Wer vorbringt, KI-Kunst sei keine Kunst,

weil sie nach keinen handwerklichen Fähigkeiten verlange, wird darauf hingewiesen, dass auch Damien Hirst seine Bilder nicht selber male und Jeff Koons seine Werke in Auftrag gebe. In Anlehnung an die Kampfparole der LGBTQ-Bewegung «Transfrauen sind Frauen», twittert ein KI-Künstler regelmässig «KI-Kunst ist Kunst» über seinen Bildern.

Nicht alle sind von der künstlichen Intelligenz so begeistert. Bald nach der Lancierung von Dall-E 2 machte der Aufruf «Fördert keinen Diebstahl!» die Runde. GPT-3, Dall-E 2 und andere KI würden Material benutzen, das unter Urheberschutz stehe, und das sei illegal. An einer anderen Stelle heisst es, Dall-E habe mit Millionen von Investorengeldern Unmengen an menschlicher Kreativität gesammelt und nicht dafür bezahlt.

Ironischerweise hat die künstliche Intelligenz ihr Metier ausgerechnet an Werken von Menschen erlernt, deren Existenz sie nun bedroht. Wer seinen Stil jahrelang perfektioniert hat, kommt sich betrogen vor, wenn jetzt andere auf derart einfache Weise die Früchte dieser Arbeit ernten können.

Da Dall-E 2 keine Plagiate erstellt, ist die Situation rechtlich Neuland. Die KI-Künstler argumentieren, was Dall-E 2 mache, unterscheide sich nicht von dem, was Künstler seit Jahrhunderten tun: inspiriert von alter Kunst neue schaffen. «Als Künstler bin ich nicht besorgt», sagt Vladimir Alexeev, der an der Universität Frankfurt Germanistik, Japanologie und Slawistik studiert hat und sich seit langem für die kreativen Möglichkeiten neuer Technologien interessiert. Er hat Kurzfilme produziert, bei denen vom Drehbuch über die Musik bis zu den Stimmen alles von künstlicher Intelligenz erzeugt wurde. Über die neuen Text-zu-Bild-Generatoren sagt er: «Ich bin inspiriert von dieser Steigerung der menschlichen Kreativität.»

«Gute Künstler kopieren. Grossartige Künstler stehlen», hat Steve Jobs oft Picasso zitiert. Doch dabei hatte er Menschen vor Augen und keine Maschine, die im Sekundentakt Meisterwerke aus den letzten 30000 Jahren ausspuckt − für 13 Cents

pro Abfrage. Nach geltendem Recht kann ein Stil nicht geschützt werden, und eine Maschine kann keine Urheberrechte in Anspruch nehmen. Das Copyright geht an die Benutzer der Text-Bild-Generatoren über.

Die offenen Fragen bei Textgeneratoren wie GPT-3 sind noch drängender. Im Juli dieses Jahres wurden an Zürcher Gymnasien Flyer mit der Frage «Wenig bis keine Zeit in deine Maturaarbeit investieren?» ausgehängt. Die Stiftung Schulwandel sucht damit nach «einem mutigen Schüler», der seine Maturaarbeit von GPT-3 schreiben lassen will. Nach gelungener Aktion soll der Scherz aufgelöst werden. Bildungsreformer wie die Stiftung Schulwandel wollen mit der KI darauf aufmerksam machen, dass in der Schule nach ihrer Ansicht das Falsche gelehrt und geprüft wird. Schon Anfang 2021 gelang es Forschern in den USA, für eine von GPT-3 in zwanzig Minuten geschriebene Seminararbeit eine genügende Note zu erhalten. Und gerade eben hat GPT-3 ihre erste wissenschaftliche Studie veröffentlicht: «Kann GPT-3 eine akademische Arbeit selbständig und mit minimalem menschlichem Aufwand verfassen?» Die Antwort heisst Ja.

Die sprachbegabten KI haben zwar keinen direkten Zugang zur Welt. Sie sind in einer riesigen fensterlosen Bibliothek gefangen, wo sie ihr ganzes Wissen aus den geschriebenen Texten beziehen, die sie während des Trainings gelesen haben. Doch das ist auch ein Merkmal der Germanistik, der Philosophie und der Geschichtswissenschaften. Alle Fächer, die sich darauf kaprizieren, aus alten Texten neue zu erstellen, werden durch die künstliche Intelligenz in Frage gestellt. Die Geisteswissenschaften wird es treffen wie einen Schlag, dass der Geist in ihrem Namen einem Rechner entsteigen kann.

«Es wird schwierig», sagt Lukas Löffel von der Universität Zürich. Löffel leitet die Abteilung Digitale Lehre und Forschung und hat auch schon mit KI rumgespielt und die Texte durch Plagiatserkennungssoftware gejagt. Erkennungsrate: null Prozent.

Im Moment vergisst GPT-3 zwar nach etwa 3000 Buchstaben, was sie zu Beginn eines Textes geschrieben hat. «Aber wir müssen uns keinen Illusionen hingeben, das geht jetzt ruckzuck», sagt Löffel. Er geht davon aus, dass es keine technische Lösung geben wird, um von einer KI erzeugte Texte zu erkennen. Die Studenten hatten zwar schon immer die Möglichkeit, einen Ghostwriter zu engagieren, aber dazu war ein gewisses Mass an krimineller Energie und Geld nötig. Mit den Sprach-KI verschwinden diese Hemmschwellen. Zudem liefert die Maschine innerhalb von Sekunden.

«Ich befürchte, dass wir an den Hochschulen zu langsam sind, um zu verstehen, was da passiert», sagt Löffel, der überzeugt ist, dass die meisten Dozentinnen und Dozenten keinen Schimmer haben. Im Herbst will er einen Workshop zum Thema anbieten. Falls er nicht weiss, was er auf die Einladung schreiben soll, weiss er, wo er Hilfe bekommt.

Natürlich gäbe es auch eine radikale Lösung für das Problem. «Die Frage ist, welche Kompetenzen ich in der Zukunft überhaupt brauchen werde», sagt Mark Cieliebak vom Zentrum für Künstliche Intelligenz an der Zürcher Hochschule für Angewandte Wissenschaften in Winterthur. «Angenommen, eine Maschine kann mir zwanzig Fachartikel in eine stimmige Arbeit verwandeln, dann muss ein Student möglicherweise nicht mehr lernen, eine Seminararbeit zu schreiben.» Cieliebak vergleicht es mit dem Taschenrechner: «Niemand lernt heute noch, Logarithmen von Hand zu berechnen.»

Doch beim Taschenrechnervergleich beschleicht viele Leute ein mulmiges Gefühl. «Natürlich erlaube ich die Verwendung von Taschenrechnern», sagt der Computerlinguist Christopher Potts von der Stanford University, «andererseits weiss ich nicht, ob Studenten, die ihre Arbeiten von einer KI schreiben lassen, dabei Fachwissen erwerben.» Selbst wenn eine KI einem bloss hilft, eine E-Mail zu

schreiben, fragt er sich: «Hat man diese Botschaft wirklich verfasst oder billigt man bloss, was jemand anderes einen drängte zu sagen?»

Die Kernfrage bleibt, wieweit die besondere Stellung der Sprache für die Art Homo sapiens beim Umgang mit der künstlichen Intelligenz eine Rolle spielen soll. Wollen wir das Finden von Argumenten, das Zusammenfassen von Information, die Genauigkeit im Ausdruck wirklich den Maschinen überlassen?

Einen originellen Umgang mit den neuen pädagogischen Möglichkeiten hat Mike Sharples vom Institut für Bildungstechnologie an der Open University in England gefunden. Er hat GPT-3 nicht nur den Auftrag gegeben, einen Essay über Lernstile zu schreiben, sondern diesen danach auch von GPT-3 bewerten lassen: *Der Aufsatz ist gut gegliedert. Er enthält eine klare Argumentation und stützt diese mit Belegen aus der Forschung. Ein möglicher Kritikpunkt ist, dass der Aufsatz nicht auf die Implikationen der Lernstilforschung eingeht oder darauf, wie diese Forschung zur Verbesserung des Lernens genutzt werden kann…*

«Die Schüler werden KI einsetzen, um Aufgaben zu schreiben. Lehrer werden KI einsetzen, um sie zu bewerten. Niemand lernt, niemand profitiert», schreibt Sharples, «wenn es jemals eine Zeit gab, die Bewertung zu überdenken, dann ist es jetzt.»

Die Unsicherheit, wie die Fortschritte in der künstlichen Intelligenz zu beurteilen sind, hat auch mit ihrer seltsamsten Eigenschaft zu tun: Niemand weiss genau, wie die Maschinen funktionieren. Das klingt paradox und ist es auch, schliesslich haben Menschen sie geschaffen, wissen genaustens über ihr Innenleben Bescheid. Wie kann es dann sein, dass die gleichen Menschen darüber rätseln, wie sie zu ihren Resultaten kommen?

Der Grund dafür liegt darin, dass sich neuronale Netzwerke grundsätzlich von herkömmlichen Computerprogrammen unterscheiden. Herkömmliche Programme

funktionieren wie Kochrezepte: Sie arbeiten eine Anweisung nach der anderen ab. Selbst wenn sie viele Millionen Zeilen lang sind und unzählige Abzweigungen enthalten, ist unser Gehirn grundsätzlich im Stand nachzuvollziehen, was Schritt für Schritt passiert. Bei einem neuronalen Netz ist das anders. Da steckt die Information nicht in Programmzeilen oder Speichern. Es gibt keine bestimmte Stelle im Netzwerk, an der die Verwandtschaft zwischen Nagel und Schraube hinterlegt ist, keinen Ort für den Unterschied zwischen «ich liebe dich» und «ich hasse dich». Vielmehr steckt dieses Wissen für Menschen nicht erkennbar verteilt in den Gewichten der Milliarden von Verbindungen.

Wenn Sie das nicht verstehen, sind Sie in guter Gesellschaft, die Experten tun es auch nicht. Sie wissen nur, dass die Netzwerke funktionieren: Nach dem Training geben sie sinnvolle Antworten. Aber auch Experten können so lange auf ein Netzwerk blicken, wie sie wollen: Sie werden weder den Nagel noch die Schraube darin finden.

Das hat eine überraschende Konsequenz: Obwohl die KI-Forscher jedes Detail ihrer neuronalen Netze kennen, bleibt ihnen nichts anderes übrig, als sie wie Ausserirdische zu behandeln. Um ihr Wesen zu ergründen, müssen sie sie verhören. Niemand weiss, welche Antworten sie liefern, bevor sie geantwortet haben.

Und so hat sich nach der Lancierung von GPT-3 und Dall-E 2 bald eine neue Tätigkeit etabliert: *die GPT-3-Therapie*. Falsch GPT-3! Die neue Tätigkeit heisst Prompt-Design. Das kann GPT-3 allerdings nicht wissen, ihre neusten Trainingsdaten stammen von 2019. Ihre Neuronen kamen weder mit Corona noch dem Ukrainekrieg in Kontakt und auch nicht mit dem Prompt-Designer, den ja erst die Arbeit an GPT-3, Dall-E 2 und den anderen KI hervorgebracht hat. Ein Prompt ist eine Anfrage an eine künstliche Intelligenz in normaler Sprache. Also: ein Wort, ein Satz, eine Frage, ein Auftrag, der Anfang eines Artikels oder einige Beispiele, die ergänzt werden sollen.

Die hochtrabende Bezeichnung Prompt-Design gilt einer auf den ersten Blick banalen Tätigkeit: Anfragen eintippen, Antworten deuten. Auf diese Weise will man

herausfinden, welche Möglichkeiten in der künstlichen Intelligenz schlummern und wie man sich mit ihr unterhält. Prompt-Designer versuchen das Potential eines hochbegabten, aber leider störrischen Wesens zu ergründen. Dazu braucht es keine Computerkenntnisse. Beantragen Sie bei OpenAI ein Passwort, und stellen Sie GPT-3 eine Frage, oder lassen Sie Dall-E 2 ein Bild erzeugen! Gratuliere, Sie sind jetzt Prompt-Designer.

Auch der Digitalkünstler Vladimir Alexeev kommuniziert auf diese Weise mit den Maschinen. Als OpenAI 2020 GPT-3 vorstellte, wurde Alexeev einer von weltweit sieben Community-Botschaftern, die den Benutzern helfen, sich zurechtzufinden. «Prompt-Design bedeutet weit mehr, als bloss Eingabetexte zu schreiben», sagt Alexeev, «man muss dafür viel Allgemeinbildung mitbringen, die Kulturgeschichte kennen und wissen, wie die Maschine tickt.»

Bei Dall-E 2 sollen schon Adjektive wie «schön» oder Verstärker wie «sehr» zu besseren Bildern führen. Jahreszahlen helfen für Bilder aus einer bestimmten Epoche. Oft wird auch der Prompt «preisgekrönt» verwendet.

Im Unterschied zur herkömmlichen Arbeit von Designern geht es bei Dall-E 2 um die Fähigkeit, Ideen in Worte zu fassen. «Ich denke, jemand wie ein Dichter könnte der beste Künstler der Zukunft sein, weil er in der Lage ist, mit der KI auf die bestmögliche Weise zu sprechen, um Dinge zu erzeugen», sagt Jordan Taylor von Vizcom, einem Start-up-Unternehmen in Kalifornien, das künstliche Intelligenz Designern und Künstlern zugänglich machen will.

Es geht auch darum, in welcher Reihenfolge eine mehrteilige Anfrage gestellt wird, welche Rolle Satzzeichen spielen oder ob die Eingabe in einer anderen Sprache erfolgen soll. GPT-3 beherrscht alle Sprachen, die in den Übungsbeispielen auftauchten. Die Schwierigkeiten bei Textaufgaben in der Mathematik liessen sich mit dem Zusatz «lass uns Schritt für Schritt vorgehen» dramatisch vermindern. Was die Arbeit eines Prompt-Designers erschwert, ist die Tatsache, dass die Maschinen auf die

gleiche Frage immer andere Antworten geben. Dall-E 2 kommt von jedem Auftrag mit vier Bildvarianten zurück, bei GPT-3 genügt ein Klick auf den Refresh-Button, um einen anderen Text zu erhalten. Diese Vielfalt ist gewollt und entsteht, weil eine Sprach-KI jeweils aus einer Palette der wahrscheinlichsten nächsten Wörter zufällig einen Begriff auslost.

Früher oder später stösst jeder Prompt-Designer auf das schmutzige Geheimnis der grossen Sprachmodelle: Sie haben die «Tendenz zu halluzinieren», wie es die Fachleute nennen. Laien würden sagen: Sie lügen, dass sich die Balken biegen. So wie Steve Jobs das Picasso-Zitat über die grossartigen Künstler, die stehlen, wohl einfach erfunden hat, so dichtet mir GPT-3 schamlos ein neues Leben an.

Die Antwort unten erhielt ich von Retobot auf die Eingabe: «Erzähle mir etwas über den besten Prompt-Designer seiner Zeit, Reto U. Schneider!»

Er wohnt in einem wunderschönen Haus am See, fährt einen flotten Wagen, hat viele Freunde und eine schöne, junge Frau. Prompt-Designer Reto U. Schneider ist der Mann fürs Leben. Aber wie ist sein Leben? Was ist aus der Wissenschaft geworden, die er anführte? Was ist aus seinen Freunden geworden? Und vor allem: Was ist aus seiner Frau geworden? Ein Liebeslied auf die Wissenschaft und ein Tribut an die Prompt-Experimente.

Eine künstliche Intelligenz beim Schwindeln zu ertappen fühlt sich an wie ein Triumph. Doch die Wahrheit zu sagen gehörte nie zu den Zielen der Maschine. GPT-3 hat kein Konzept davon, was richtig oder falsch ist. Bei aller Faszination für die Antworten muss man sich immer wieder vergegenwärtigen, dass ihre erstaunliche Leistung dem simplen Vorgang entspringt, das wahrscheinlichste nächste Wort zu erraten. Selbstzweifel entwickelt die künstliche Intelligenz auf diese Weise keine. Ihrem Einsatz als Schreibassistent tut das keinen Abbruch. Wenn man die Texte prüft, kann die Arbeitsersparnis beim Erstellen von Blogbeiträgen oder Social-Media-Posts enorm sein.

Natürlich wäre es den Entwicklern lieber, ihre Sprachmaschinen gäben von Anfang an richtige Antworten und würden zugeben, wenn sie etwas nicht wüssten. Aber wie man das hinbekommt, ist unklar. Ein anderer Makel ist genauso schlimm: Weil die Maschinen an Beispielen aus dem Internet trainiert sind, transportieren sie das Weltbild, das diesen Trainingsdaten zugrunde liegt, ganz egal, wie sexistisch oder rassistisch es ist. Die Daten von Hand auszuwählen ist wegen ihrer enormen Menge nicht möglich. Man kann zwar kuratierten, sauberen Übungsbeispielen beim Training mehr Gewicht verleihen, aber eine perfekte Lösung ist das nicht. Deshalb entlassen die Firmen ihre Kreaturen nur mit Maulkorb in die freie Wildbahn.

Sowohl GPT-3 wie auch Dall-E 2 haben Filter vorgeschaltet, die gewisse Begriffe abfangen und vulgäre oder sexuelle Aufträge verweigern. GPT-3 erteilt keine rechtlichen oder medizinischen Ratschläge. Bei Dall-E 2 werden keine Gesichter lebender Personen dargestellt – und auch jene von Mohammed und Gott nicht. Gegen das Verbreiten von Klischees hilft das allerdings nicht.

Wer nach einem CEO verlangt, bekommt Bilder von Männern präsentiert, Pflegepersonal sind immer Frauen, und Hochzeitsbilder zeigen heterosexuelle Paare. Im Juli 2022 gab OpenAI bekannt, man habe nun eine neue Technik eingeführt, «damit Dall-E Bilder von Menschen erzeugt, die die Vielfalt der Weltbevölkerung besser widerspiegeln». Die neuen Bilder zeigten tatsächlich weniger Stereotype, und die Benutzer fragten sich, wie OpenAI das so schnell hinbekommen habe, denn das Nachtrainieren mit neuen Bildbeispielen ist aufwendig.

Tests legten schliesslich nahe, dass Dall-E 2 wahrscheinlich nicht neu trainiert worden war. Vielmehr wurden die Anfragen wohl für die Benutzer unsichtbar mit weiteren Begriffen wie «Frau» oder «asiatisch» versehen, um die Diversität zu erzwingen.

Das ist eine kümmerliche Lösung für ein viel grundsätzlicheres Problem. Die Menschen sind aus nachvollziehbaren Gründen unterschiedlicher Meinung. Wer darf

entscheiden, welche Werte eine künstliche Intelligenz verbreitet? Bis jetzt sind es die privaten Betreiber der Programme. Und weil sie in ständiger Angst vor Shitstorms und rechtlichen Problemen leben, sperren sie lieber zu viel als zu wenig. So gehören zum Beispiel «Ukraine» und «Russland» zu den verbotenen Begriffen bei Dall-E 2. Und die restriktive Haltung gegenüber Gewaltbegriffen verbietet der künstlichen Intelligenz, eine Stelle aus Gotthelfs Novelle «Die schwarze Spinne» zu illustrieren.

So viel Kontrolle in der Hand gewinnorientierter Unternehmen liess Alternativen entstehen. Der frühere Hedge-Fund-Manager Emad Mostaque hält es für paternalistisch, dass uns Firmen vorschreiben, wie wir ihre künstliche Intelligenz verwenden dürfen. Er gründete die Organisation StabilityAI mit dem Motto «AI from the people for the people», die Experten und Laien einen Zugang zu grossen KI-Maschinen mit weniger Einschränkungen ermöglichen soll. Doch auch er musste harsche Kritik aus dem Volk einstecken. StabilityAI sei für Illustratoren und Künstler keine geringere Bedrohung als Dall-E 2. Zudem werde die laschere Kontrolle zwangsläufig zu Auswüchsen führen. Soll eine künstliche Intelligenz die Realität abbilden oder die Welt, die wir uns wünschen?

Der Geist ist aus der Flasche. Wöchentlich wird eine neue Maschine angekündigt, jede mit noch mehr Neuronen, die an noch mehr Übungsbeispielen trainiert wurden als die vorangegangene. Mittlerweile gibt es auch Gratisangebote wie die KI Craiyon.

Wo das hinführt, mag niemand prophezeien, doch eine faszinierende Entwicklung ist absehbar: Wie auf die Fotografie der Film folgte, werden auch die KI-Bilder in Bewegung geraten. Dann wird ein Regisseur sich vor einen Computer setzen und seinen nächsten Kino-Hit diktieren. Und wir *werden wieder staunen, wie selbständig sich die Bilder den Regeln unserer Welt angepasst haben. Vielleicht werden wir dabei an John Lennons Bemerkung über den Rock nach dem Beatles-Hype denken: «Alles ist jetzt Rap.» Alles ist jetzt KI.* – Oder hat Lennon das gar nie gesagt?