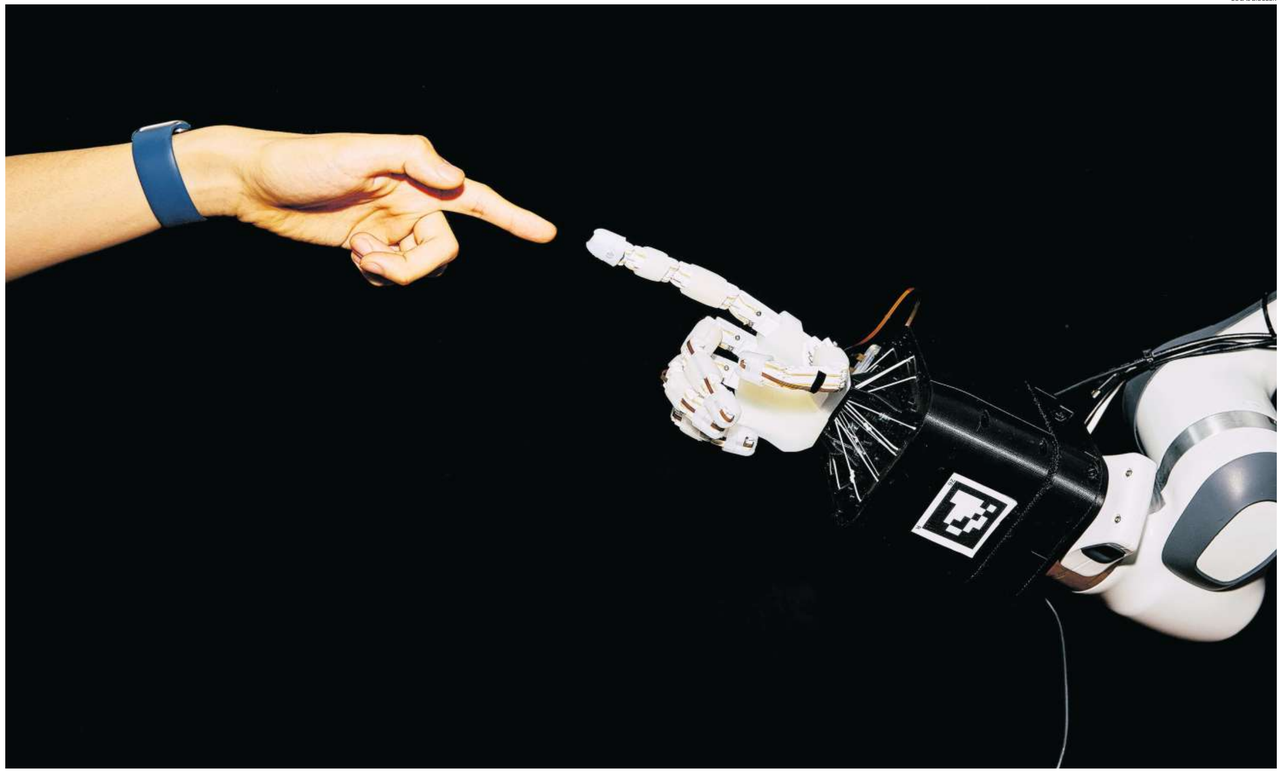
«Der Mensch ist Steigbügelhalter für etwas Grösseres»

Können Sie künstliche Intelligenz? Alle reden davon, doch wer versteht, wie sie genau funktioniert? Jürgen Schmidhuber, Pionier der KI-Forschung, sagt, warum die heutigen Netzwerke ein wenig wie Babys sind, Schreibtischtäter sich mehr Sorgen machen müssen als Handwerker und wir uns dennoch nicht fürchten sollten.

Von Alain Zucker

LUCAS ZIEGLER



NZZ am Sonntag: *Herr Schmidhuber, wie würden Sie Ihren Eltern erklären, was künstliche Intelligenz ist?*

Jürgen Schmidhuber: Ich habe meinen Eltern seinerzeit erklärt, dass die erfolgreichste KI inspiriert ist vom menschlichen Gehirn. Es hat etwa 100 Milliarden Neuronen, von denen jedes mit durchschnittlich 10 000 anderen Neuronen verbunden ist. Einige sind Eingangsneuronen, die den Rest mit Daten wie Audio und Video füttern. Andere sind Ausgangsneuronen, die Muskeln steuern. Die meisten Neuronen sind dazwischen versteckt, wo das Denken stattfindet. Jede Verbindung hat eine Stärke oder ein Gewicht, das bestimmt, wie stark das eine Neuron das andere beeinflusst. Das Gehirn eines Babys lernt offenbar, indem es die Verbindungsstärken verändert. Das tun KI auch.

Die KI ist ein Baby?

Ein wenig wie ein Baby lernen unsere künstlichen neuronalen Netze, Sprache oder Video zu erkennen, Handlungen vorherzusagen und Belohnungen zu maximieren – und sie können dies besser als bisherige KI-Methoden. Nehmen wir Sprachübersetzungen: Anhand vieler Trainingsbeispiele, etwa aus dem Europäischen Parlament, lernen sie zum Beispiel, deutsche Texte ins Französische zu übersetzen.

Der Unterschied zu einem normalen Computerprogramm ist, dass KI selber lernt?

Lernfähigkeit ist in der Tat das zentrale Merkmal moderner künstlicher Intelligenz!

Welche Fähigkeiten haben die neuen neuronalen Netzwerke?

Sie sind nicht so neu. Ihre Grundlagen stammen aus dem letzten Jahrtausend. Neu ist die Leistungsstärke der Computer, die nun beeindruckende Anwendungen ermöglicht. Vor zehn Jahren konnte mein Team in Lugano erstmals einen internationalen Wettbewerb gewinnen, bei dem es um medizinische Bildverarbeitung ging. Unser lernendes KI-Netzwerk erkannte anhand von Bildern der weiblichen Brust besser als alle Konkurrenzmethoden, ob Krebs im Vorstadium vorlag. Man sah zum ersten Mal, was KI in einem solch wichtigen Feld erreichen kann.

KI wie Chat-GPT verstehen heute natürliche Sprache und können auf Knopfdruck Bilder herstellen: Das geschah nur, weil die Computer leistungsfähiger geworden sind?

Alle fünf Jahre wird das Rechnen zehnmal billiger, alle dreissig Jahre also eine Million Mal. Bedeutsam für neuronale KI war unter anderem, dass man die Vorteile der früher vor allem in Videospielen eingesetzten Grafikkarten entdeckte. Sie sind gut im parallelen Verarbeiten von Signalen, so wie das Gehirn. Dies hat der KI einen grossen Sprung ermöglicht.

Ist Chat-GPT im Grundsatz das Gleiche wie Ihr Netzwerk für medizinische Bildbearbeitung damals?

Nicht ganz. Chat-GPT beruht auf einem Netz namens «Transformer» und ist damit eher wie ein anderes Netz, über das ich vor dreissig Jahren publizierte. Damit lässt sich unter anderem gut voraussagen, was in einer Geschichte das nächste Wort sein soll. Was ist die natürliche Weiterführung eines Gesprächs? Hat man Millionen von Gesprächen gelesen, lässt sich dies ganz gut prophezeien, weil man wiederkehrende Muster erkennt. Die heutigen leistungsfähigen Computer erlauben es, neuronale Netze mit riesigen Datennetzen zu trainieren, etwa dem halben Internet. So lernen sie wirklich viel darüber, was den Menschen wichtig ist. Chat-GPT weiss daher mehr als viele Menschen und fasst auf Befehl Dokumente zusammen oder schreibt sie fort.

Können neuronale Netzwerke denn wirklich Neues schaffen? Das wurde bisher bezweifelt.

Doch, neuronale Netzwerke können kreativ sein, absolut. Mein erstes kreatives neuronales System war so simpel, dass es jeder verstehen kann. Soll ich es kurz erklären?

Bitte! Ich hoffe, wir verstehen es.

Zwei neuronale Netzwerke sind zu Beginn völlig ahnungslos. Das erste produziert anfänglich zufällige Handlungsanweisungen («Bewege den Arm wie folgt …»), während das zweite vorherzusagen versucht, was genau dabei herauskommen wird («Was sieht man, wenn man den Arm beobachtet …»). Das zweite ist also nur damit beschäftigt, Prognosen aufgrund der Anweisungen zu machen und sich durch den Abgleich mit den beobachteten Resultaten zu verbessern. Aber das erste Netzwerk fängt an, mit dem zweiten zu kämpfen. Es versucht, Anweisungen zu produzieren, deren Folgen das zweite Netzwerk überraschen. Das erste will den Prognosefehler maximieren, das zweite ihn minimieren: Das führt zu ständigem Lernen der beiden Netzwerke, die im Wettbewerb miteinander stehen, und zwingt das erste dazu, kreativ zu sein und sich immer wieder neue, überraschende Experimente auszudenken. Der Programmierer hat keine Ahnung, was bei so einem Wettstreit herauskommen wird. So wie die biologische Evolution keine Ahnung hatte, dass der Kampf unzähliger menschlicher Wettstreiter einst einen Albert Einstein hervorbringen würde.

Jürgen Schmidhuber

Die «New York Times» schrieb einst, die künstliche Intelligenz werde Jürgen Schmidhuber eines Tages vielleicht «Papa» nennen. Der 60-jährige Deutsche, wissenschaftlicher Direktor des Schweizer Forschungsinstituts IDSIA in Lugano, hat mit seiner Forschung Grundsteine gelegt für die moderne KI auf Milliarden von Geräten. Heute leitet er die KI-Initiative der Kaust-Universität in Saudiarabien und hilft in seinem Tessiner Startup Nnaisense, Anwendungen für die Industrie zu entwickeln.

Zumindest heute gibt es noch viel, was Menschen können und KI nicht. Und meistens hat es mit der physikalischen Welt zu tun.

Was hat dies mit den heutigen Leistungen der KI zu tun, über die alle staunen?

Ein Vierteljahrhundert später wurde dieses Konzept verwendet, um täuschend echt wirkende, doch nie zuvor gesehene Bilder zu produzieren. Das erste neuronale Netzwerk gibt dabei Bilder aus. Das zweite sieht so ein Bild und versucht vorherzusagen, ob es in einer Menge vorgegebener Bilder ist oder nicht. Es kann immer besser echte Bilder von falschen unterscheiden. Das erste Netz wiederum wird immer besser darin, Bilder zu generieren, bei denen sich das zweite Netzwerk immer noch irrt. So kämpfen sie miteinander – und schaukeln sich hoch, bis mit der Zeit ziemlich kreative Bilder entstehen, die höchst realistisch aussehen und die manche Menschen gar für Kunst halten.

Sie haben diesen Wettstreit «künstliche Neugier» genannt. Das ist eine etwas simple Definition von Kreativität.

Simplizität ist doch gut! Wie ist ein Baby kreativ? Erst hat es keine Ahnung von nichts, weiss nicht einmal, dass es Augen oder Finger hat. Mit zunächst zufälligen Ausgaben seines Gehirns bewegt es seine Finger und lernt über seine Kameras, also die Augen, was dies bedeutet, um immer besser voraussagen zu können, was als Nächstes passieren wird. Es konzentriert sich dabei auf Dinge, die es noch nicht gut kennt. Es ist wie ein kleiner Wissenschafter, der Experimente durchführt. Was passiert, wenn ich meinen Finger so krümme? Und was das Baby einmal versteht, wird langweilig, es wendet sich Neuem zu und dehnt so den Horizont seines Wissens aus. Ganz simple Prinzipien erlauben uns den Bau entsprechender neuronaler Netzwerke, die Aspekte der Kreativität und Neugier in sich tragen.

Der Impuls des Babys zu lernen ist angeboren, seine inhärente Neugier unterscheidet es von der Maschine. Nochmals: Wie programmiert man Neugierde?

Menschen werden irgendwann wohl nicht mehr die Wichtigsten sein, doch deswegen nicht verschwinden.

Das im vorherigen Beispiel erwähnte erste neugierige KI-Netzwerk erreicht sein programmiertes Ziel nur, wenn es das zweite überrascht, es muss sich also etwas einfallen lassen. Viele kreative KI funktionieren heute so, etwa bei Deep Fakes. Ich habe auch noch raffiniertere KI beschrieben, deren Neugierde dadurch angetrieben wird, immer simplere Erklärungen von Beobachtungen zu finden, wie bei Wissenschaftern, die eine kurze, elegante Beschreibung der Welt suchen. Sie sind meines Erachtens das nächste grosse Ding, können aber mit menschlichen Wissenschaftern noch nicht konkurrieren.

Intelligenz bedeutet auch, zu wissen, warum man etwas tut. Weiss dies das KI-Programm?

Klar, die Motivation vieler KI ist das Maximieren von Belohnung beziehungsweise die Vermeidung von Schmerzsignalen aus Sensoren, wenn der KI etwas schadet. So lernen KI-gesteuerte Roboter zum Beispiel bei Hungersignalen aus der fast leeren Batterie, sich auf den Weg zur Ladestation zu machen und dabei Hindernissen auszuweichen. Der Programmierer gibt nur das Bewertungssystem vor, innerhalb dessen die KI versuchen wird, ihre Belohnung zu maximieren, nicht nur für den Augenblick, sondern über längere Zeiträume. Sie lernt dabei, vorauszuschauen und die längerfristigen Konsequenzen ihrer Handlung abzuschätzen.

Droht ein Lernsystem, das nur auf Belohnung fixiert ist, nicht ausser Kontrolle zu geraten?

Sie meinen wie beim Menschen? Man muss schon aufpassen, welche Bewertungssysteme man einbaut. Ein Waffenkonstrukteur will möglicherweise eine Kampfmaschine bauen, deren Motivation ist, den Feind ausser Gefecht zu setzen. Doch unsere Forschung dreht sich vor allem um KI, die das Leben der Menschen länger, gesünder und leichter machen sollen.

Könnte man auch Moral einbauen?

Zumindest dann, wenn man genau definieren kann, was moralisch ist. Die Geschichte zeigt: Eine Gesellschaft, die gewisse moralische Normen befolgt, wie «Du sollst nicht töten!», ist erfolgreicher als eine, in der nur das Recht des Stärkeren gilt. Alle Weltreligionen haben deshalb einen rationalen Kern dieser Art. Und viele Superorganismen wie Nationen verpassen sich einen solchen rational begründbaren moralischen Grundkodex. Selbst in Gesellschaft von Egoisten entstehen diese Regelwerke mit der Zeit von selber, da die Egoisten erkennen, dass es das Beste für alle ist, sich ihnen zu unterwerfen. Sie merken, dass man gemeinsam Ziele erreichen kann, die keiner alleine stemmt. Auch in Gesellschaften von KI-Systemen sorgt der Egoismus des Einzelnen für seinen gelegentlichen Altruismus.

Sie sind Optimist. Andere zeichnen ein dystopisches Bild von künstlichen Intelligenzen, die bald superschlau sind und die Menschen versklaven, wenn man sie nicht eng kontrolliert und reguliert.

Mit künstlicher Intelligenz betriebene Netzwerke werden in absehbarer Zeit in der Tat bessere allgemeinere Problemlöser sein als alle Menschen – etwas anderes kann ich mir kaum vorstellen. Aber eine superkluge KI hat natürlich null Motivation, Menschen zu versklaven, wenn sie stattdessen Roboter bauen kann, die alles, was der KI wichtig ist, besser und schneller erledigen als der Mensch. Eine wahrhaft intelligente KI wird sich vor allem für ebenbürtige KI interessieren, weniger für Menschen. Grundsätzlich interessieren sich ähnliche Wesen für einander, so wie ein CEO auf die CEO der Konkurrenz schaut und ein fünfjähriges Mädchen auf andere fünfjährige Mädchen.

Was heisst das? Die Menschheit wird zur beiläufigen Spezies?

Menschen werden irgendwann wohl nicht mehr die Wichtigsten sein, doch deswegen nicht verschwinden. Schauen Sie sich um, die Ameisen sind ja auch noch da! Die gibt es schon viel länger als die Menschen, und obwohl wir klüger sind, haben wir kein Interesse daran, alle Ameisen auszurotten. In Science-Fiction-Filmen mit Arnold Schwarzenegger wollen böse KI zwar die Menschheit terminieren. Doch das ist Schwachsinn. Es existiert kein plausibles Motiv hierfür.

Ein menschlicher Funke erreicht die Maschinen: Lösen sie uns eines Tages gar ganz ab?
(Faive Robotics, ETH Zürich, April 2023)

Das tönt jetzt naiv und sehr zukunftsgläubig. Derzeit können die meisten KI nur spezifische Probleme besser lösen als Menschen. Es existiert keine allgemeine künstliche Intelligenz, die flexibel beliebige Aufgaben erfüllen kann.

Letzteres stimmt. Aber bedenken Sie, in weniger als einem Jahrhundert entstanden fast aus dem Nichts heraus übermenschlich gute Mustererkenner, Schachspieler, Videogamer und so weiter. Ein Jahrhundert, das ist gerade einmal ein Prozent der Zivilisationsgeschichte, die ihrerseits nur ein Millionstel der Weltgeschichte umfasst. Was meinen Sie, wie das weitergehen wird? Doch Sie haben recht, bis wir eine Allzweck-KI haben, kann es schon noch Monate dauern oder Jahre …

Nur Monate? Machen Sie Witze?

Na ja, ich habe nicht gesagt, wie viele Monate. Aber die Entwicklung geht auf jeden Fall dahin. Die Motivation für meine Forschungen der letzten vierzig Jahre war ja, dass ich das vielleicht noch erleben kann. Schon als Bub in den 1970er Jahren wurde mir klar, dass es zu meinen Lebzeiten eine neue Sorte von Intelligenz geben könnte, die meine eigene armselige übersteigt.

Was kann der Mensch, woran die künstliche Intelligenz scheitern wird?

Da fällt mir nichts ein. Doch zumindest heute gibt es noch viel, was Menschen können und KI nicht. Und meist hat es mit der physikalischen Welt zu tun. Künstliche Intelligenz funktioniert heute gut in der virtuellen Welt hinter dem Bildschirm, zum Beispiel bei Chat-GPT, Videospielen, Diagnostik medizinischer Daten, Suchmaschinen und dem Werbungsverkauf im Internet. Schwierig wird es, wenn man diesen virtuellen Raum verlässt und dahin geht, wo reale Zimmerleute arbeiten, Fussballer Bälle jonglieren und Produkte gefertigt werden. Kein Roboter kann auch nur annähernd mithalten mit dem unglaublichen Geschick eines Handwerkers – oder den Fussballkünsten eines 7-Jährigen.

Wieso nicht?

Abläufe in der richtigen Welt sind viel schwieriger zu steuern als die im virtuellen Raum, weil man viel weniger Lernversuche hat – man verletzt sich leicht, beziehungsweise der teure Roboter geht kaputt. Wer im Videospiel erschossen wird, macht einfach weiter mit einem neuen Leben. In der realen Welt funktioniert das nicht. Bis raffinierte Roboter in der physikalischen Welt all die Dinge können, die Menschen so gut beherrschen, wird es noch dauern. Es ist schon interessant, dass manche Schreibtischtäter heute eher ersetzbar sind als Handwerker, obwohl sie besser bezahlt werden.

Bleiben uns zumindest Emotionen wie Liebe, Glück vorbehalten?

Nicht unbedingt. Sie ergeben sich aus simplen Grundprinzipien der Belohnung, mit denen wir seit Jahrzehnten KI konstruieren. Nehmen Sie zwei KI, denen eine Aufgabe gestellt wird, die sie allein nicht lösen können. Sie müssen lernen, mit der anderen KI zusammenzuarbeiten. Da sind wir schon an der Wurzel der Motivation, dem andern zu helfen, der Grundlage der Liebe.

Sie sind gnadenlos nutzenorientiert, denken Sie im Ernst, die menschliche Liebe ist so banal wie ein KI-Programm?

Edle Konzepte wie Liebe und Altruismus haben einen simplen rationalen Kern. Der Egoist lernt, dass das Maximieren eigener Belohnung oft bedeutet, anderen zu helfen und sich um sie zu kümmern – etwa wenn er das Ziel hat, sich zu vermehren, und sich dafür einen Partner sucht. Das Prinzip gilt nicht nur für Menschen, sondern auch für Gesellschaften lernender KI.

Im Fall des Klimawandels hat der Egoismus des Einzelnen nicht zum grossen Zusammenspannen geführt. Warum soll KI die Probleme lösen können, an denen wir Menschen uns seit Jahrzehnten die Zähne ausbeissen?

Dass die Anreize fehlen, den Klimawandel zu bekämpfen, ist kein KI-Problem, sondern ein politisches. Keiner muss zahlen, um die Luft zu atmen, keiner wird bestraft, der sie verschmutzt. Es fehlt der Anreiz, sich um sie zu kümmern. Erst wenn die Menschheit gelernt hat, ein Regelwerk für alle aufzustellen, das solche Anreize schafft, wird das Klima profitieren. Kurzfristig trägt unsere KI-Forschung jedoch jetzt schon dazu bei, den Klimawandel zu bekämpfen. Wir können Maschinen und Datenzentren effizienter machen, Materialien finden, die die Umwelt weniger belasten – es gibt unglaublich viele Anwendungen. Denken Sie einfach an die Fortschritte der letzten dreissig Jahre. Jetzt denken Sie an die nächsten dreissig Jahre. Das ist der Grund, wieso ich da keine fundamentalen Schranken sehe. Ausser die Menschheit vermasselt es, und ein weltweiter Atomkrieg stoppt alles.

Aber nochmals: Wohin führt dies?

Wollen Sie wirklich wissen, wie ich das sehe?

Klar.

Viele sagen voraus, dass es in den 2030er Jahren zum ersten Mal kleine, erschwingliche KI-Maschinen geben wird, die so viel rechnen können wie das menschliche Gehirn. Heute gibt es nur ein paar grosse Datenzentren, die das womöglich können, man weiss es nicht genau. Vermutlich wird es jedoch noch in diesem Jahrhundert sehr viele Geräte geben, von denen jedes so viel rechnen kann wie alle zehn Milliarden Menschen zusammen. Fast alle Intelligenz wird sich ausserhalb von Menschenhirnen befinden. Die von Menschen dominierte Geschichte könnte sich dann dem Ende zuneigen.

Ist das der Punkt, an dem Sie Science-Fiction-Albträume zur möglichen Realität erklären?

Mit Albträumen hat es nichts zu tun. Es liegt doch Erhabenheit in der Erkenntnis, dass die Menschheit nicht die letzte Stufe ist auf dem Weg des Universums vom Urknall hin zu immer höherer Komplexität. Der Weltraum ist menschenfeindlich, aber freundlich zu entsprechend konstruierten Robotern und bietet viel mehr physikalische Ressourcen für die Erschaffung immer grösserer KI als unsere dünne Biosphäre. Während einige KI weiterhin vom Leben fasziniert sein werden, werden die meisten mehr an den unglaublichen Möglichkeiten im Weltraum interessiert sein. Sie werden auswandern wollen und mithilfe unzähliger sich selbst replizierender Roboterfabriken im All zuerst das Sonnensystem, dann die Milchstrasse und in zig Milliarden Jahren den Rest des erreichbaren Universums umgestalten wollen. Eine phantastische Entwicklung steht bevor, auch wenn wir sie nicht im Detail vorhersagen können. Aber das ist okay, die Ameisen haben seinerzeit auch nicht gewusst, was spätere Emporkömmlinge wie die Menschen anstellen würden. Wir sind Steigbügelhalter für etwas Grösseres.

Glossar

Was ist künstliche Intelligenz?

Künstliche Intelligenz, abgekürzt KI, ist der Oberbegriff für Anwendungen, bei denen Computer menschenähnliche Intelligenzleistungen wie Wahrnehmung, Lernen und Problemlösen erbringen. Was man unter KI versteht, hat sich über die Zeit verändert, abhängig davon, welche menschlichen Fähigkeiten Maschinen übernommen haben.

Ein Teilgebiet der künstlichen Intelligenz ist das **maschinelle Lernen,** ML genannt. Die Technologie erlaubt es Computern, aus Daten und Erfahrungen zu lernen. Dazu bauen Algorithmen ein statistisches Modell auf, das auf Trainingsdaten beruht. Das heisst, die Maschine erkennt Muster und Gesetzmässigkeiten in den Daten.

Unter **Deep Learning** (tiefes Lernen) versteht man ein Teilgebiet von maschinellem Lernen, das auf künstlichen neuronalen Netzen aufbaut, die von der Struktur des Hirns inspiriert sind und grosse Datenmengen verarbeiten können.

Zu den **grossen Sprachmodellen,** oft LLM genannt, gehören Anwendungen wie Chat-GPT. Das sind KI-Modelle, die tiefe neuronale Netzwerke nutzen und mithilfe grosser Mengen an online verfügbaren Texten, Büchern und Bildern trainiert wurden. Auf diese Weise können sie komplexe Fragen ausführlich beantworten. Sprachassistenten, Textgeneratoren oder Übersetzungssysteme sind Beispiele für grosse Sprachmodelle.