Wer knackt den Hirncode?

Bestimmte einfache Testaufgaben überfordern die künstliche Intelligenz bis heute. Erst wenn sie diese lösen kann, wird sie vergleichbar mit der Intelligenz von Menschen. Was sagt das aus über das Wesen der natürlichen Intelligenz?

Von Eva Wolfangel (Text) und Matthieu Bourel (Illustration), 10.07.2023

Wie kann man eine Maschine dazu bringen, diese Aufgaben zu lösen? Es geht um das Herumschieben bunter Quadrate, und die sind überall zu sehen in der altehrwürdigen Villa in Davos. Hier ist ein ungewöhnliches Labor untergebracht: Im Lab 42 gibt es keine Reagenzgläser, keine Laborkittel und keine brummenden Kühlschränke. Bloss Computer.

Der Name des Lab ist Informatikerhumor und bezieht sich auf den Roman «Per Anhalter durch die Galaxis». Darin ist «42» die Antwort auf alles. Und offenbar führt für die Antwort auf alles kein Weg vorbei an den bunten Quadraten. Sie werden in der Villa an die Wand projiziert, auf Plakaten und

in Präsentationen gezeigt, leuchten auf Bildschirmen. Mehrere Quadrate bilden zusammen jeweils ein Muster. Vorgegeben sind zwei dieser Muster, A und B. Aus dem Unterschied zwischen A und B lässt sich ablesen, nach welchen Regeln das Muster verändert werden muss, um A in B umzuwandeln. Beim dritten Beispiel ist nur A vorgegeben – und die Aufgabe ist es, das Muster B zu generieren.

Für Menschen ist das kein Problem (verschaffen Sie sich ein Erfolgserlebnis und klicken Sie sich durch ein paar Aufgaben). Jedes Schulkind kann die ersten Aufgaben des sogenannten Abstraction and Reasoning Corpus (ARC) lösen. Und mit diesem Modell soll sich nun eine künstliche Intelligenz, kurz: KI, entwickeln lassen, die uns Menschen gleichkommt oder gar übertrifft? Genau dies ist die Hoffnung im Lab 42.

Das Einfachste ist am schwersten

Die Idee: Wenn Maschinen diese Aufgaben lösen können, haben sie eine entscheidende Fähigkeit erworben – nämlich zu abstrahieren und zu generalisieren. Weltweit tüfteln deshalb Tausende Fachleute an Computerprogrammen, die 400 solcher ARC-Aufgaben lösen sollen. Die Aufgaben sind jedoch eine der wenigen Herausforderungen, an denen Computer bisher grandios scheitern. Trotz KI-Hype um Chat GPT und andere künstliche Sprachsysteme, die immer schlauer zu werden scheinen.

Während Unternehmer wie Elon Musk, der Historiker und Bestsellerautor Yuval Noah Harari oder Apple-Gründer Steve Wozniak davor warnen, dass KI die Weltherrschaft übernehmen und zu Massenarbeitslosigkeit führen könnte, und andere davon schwärmen, dass KI die Klimakatastrophe und den Ernährungsnotstand lösen werde, sitzt Michael Hodel an diesem Sommertag im Lab 42 in Davos und fragt sich, wie er KI befähigen kann, diese denkbar simplen Aufgaben zu lösen.

Der 26-jährige Informatikstudent der ETH Zürich – lange Haare, Sporthosen und verwaschenes T-Shirt – versucht gerade, weitere ARC-Aufgaben zu kreieren, damit ein Computerprogramm daraus Konzepte lernen kann. Als Informatiker geht Hodel das Ganze logisch an, und das heisst für ihn, dass die ARC-Aufgaben auf Konzepten beruhen. Beispielsweise auf dem Konzept, einen Hohlraum zu füllen. Oder zwei Objekte aufeinanderzustapeln. «Bisher hat das niemand mit maschinellem Lernen gelöst», sagt Hodel. Das liegt daran, dass KI sehr viele Trainingsdaten braucht – die 400 Aufgaben sind längst nicht genug.

Im Lab ist allerhand los an diesem Tag: Sechs der acht Mitarbeiterinnen treffen sich zur monatlichen Klausur, zudem kommt ein Investor, der sich persönlich von den Fortschritten überzeugen will, die das Lab auf dem Weg zur sogenannten «Allgemeinen Künstlichen Intelligenz» macht, auf Englisch: Artificial General Intelligence (AGI). Das ist der Fachausdruck für das, was sich das Lab 42 als Ziel vorgenommen hat: die Entwicklung einer künstlichen Intelligenz, welche die menschlichen Fähigkeiten übertrumpft.

Heute ist KI noch eine Fachidiotin

«Decode the mind for humankind» prangt in grossen Buchstaben auf der Website des Lab 42. «Wir wollen den Braincode entschlüsseln», sagt der Zürcher Neurowissenschaftler und Investor Pascal Kaufmann stolz. Er hat das Lab 42 gegründet und beschäftigt sich seit einigen Jahren mit der Frage, was sich Computer vom menschlichen Gehirn abschauen können, um ihre «Intelligenz» auf breitere Füsse zu stellen. Denn bisher sind maschi-

nelle Systeme zumeist eine Art Fachidiot: Sie können in der Regel nur eine einzelne Aufgabe sehr gut lösen. Beispielsweise Bilder erkennen, die beste Route von A nach B errechnen oder menschliche Sprache erzeugen. Neueste Sprachmodelle wie Chat GPT wirken zwar aufgrund ihrer Eloquenz erstaunlich intelligent – aber immer wieder fallen sie auch durch dumme Fehler auf, die zeigen, dass sie nicht «verstehen», wovon sie sprechen.

Wer die Tür gegenüber von Michael Hodels Arbeitsplatz öffnet, kommt, so steht es auf dem Türschild, in den «Boardroom». Und taucht in einen blauen Lichtschein, der von einer grossen, leuchtenden Nachbildung des menschlichen Gehirns ausgeht. Die leuchtende Plastik nimmt den ganzen Erker ein und beleuchtet abends durch die Fenster auch die Strassen von Davos. «Damit alle wissen: Das sind die mit dem Gehirn», sagt Kaufmann grinsend. In seiner weissen Jacke mit den «Lab 42»-Aufnähern an der Schulter erinnerte er an einen Nasa-Astronauten, als sich auf dem Parkplatz vor dem Lab die Flügeltüren seines weissen Tesla hoben und er ausstieg wie aus einem Raumschiff.

Kaufmann setzt sich an den grossen runden Tisch, der den Boardroom dominiert, und versucht, seinen Laptop mit dem Beamer zu verbinden. Gleich kommt der Investor, dem er von den neuesten Projekten berichten will. Als der Beamer läuft, erscheinen auch hier die bunten Quadrate an der Wand.

Die Fragen, die Kaufmann umtreiben und zur Gründung des Lab 42 führten: Wie könnte eine künstliche Intelligenz entwickelt werden, die flexibel ist wie die menschliche Intelligenz? Die gleichermassen Aufgaben aus verschiedenen Bereichen lösen kann? Muss sie dafür abstrakte Zusammenhänge verstehen lernen? Und falls ja: Was bedeutet Verstehen eigentlich?

Der Weg, um diese Fragen zu beantworten, könnte über die Lösung eines Wettbewerbs führen, den Kaufmann mit seinem Lab ausgeschrieben hat. Wer ein Computerprogramm einreicht, das alle ARC-Aufgaben lösen kann, gewinnt 69'000 Franken. Tausende Teams aus aller Welt haben sich bereits am Wettbewerb beteiligt. Die gesamte Challenge besteht aus 400 Beispielaufgaben, auf deren Basis Maschinen lernen sollen. Und 400 weiteren Testaufgaben, die Forschende nutzen können, um zu testen, ob ihre Computersysteme in der Lage sind, Aufgaben, mit denen sie nicht trainiert worden sind, mit demselben Programm zu lösen. Und dann gibt es noch 200 geheime Testaufgaben, die nur einer kennt: François Chollet, ein Google-Forscher, der die Aufgaben entwickelt hat und damit die eingereichten Lösungen bewertet.

Bisher haben die meisten Teilnehmenden null Prozent erreicht. Nur eine Handvoll Teams konnte mehr als 10 Prozent der Aufgaben lösen, der aktuelle Weltrekord liegt bei 31,4 Prozent. Für jeden weiteren Prozentpunkt verspricht Kaufmann 1000 Franken – so kommen die insgesamt 69'000-Franken zusammen. Und die ersten drei Teams, die 42 Prozent erreichen, können ihren Namen in einen Felsen in den Schweizer Alpen eingravieren lassen.

Student Hodel ist gewissermassen Kaufmanns Trumpf. Bis vor kurzem war er Weltrekordhalter: Hodel hat ein Programm entwickelt, das 30,4 Prozent der 400 Testaufgaben gelöst hat. Er kam über eine *summer school* ins Lab, zu der Kaufmann geladen hatte. Obwohl Hodel kein Mitarbeiter des Lab ist, darf er trotzdem kommen, wann immer er will – für ihn ist immer ein Computer frei. Schliesslich hatte Kaufmann seinen Investoren schon für 2022 einen Weltrekord versprochen, den Hodel mit etwas Verspätung Anfang-2023 schliesslich geschafft hatte. Und auch wenn Hodels Bestmarke kürzlich übertroffen worden ist, bleibt er weltweit einer der wenigen Program-

mierer, die überhaupt einige der Aufgaben maschinell lösen lassen konnten. Und, vor allem: die ein Gefühl dafür haben, welche Methoden funktionieren könnten.

Was ist denn eigentlich für Maschinen so schwer an dieser Challenge? Und welche entscheidende Fähigkeit sollen sie daraus lernen, die sie näher an allgemeine Intelligenz bringt?

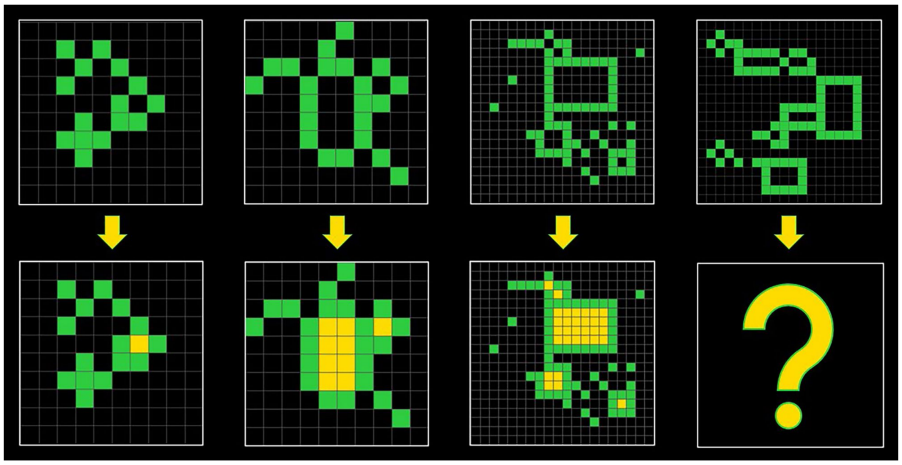
Was Intelligenz nicht ist

Es lohnt sich, dazu den Erfinder des Abstraction and Reasoning Corpus zu befragen. François Chollet erklärt sich sofort bereit, über sein Herzblutprojekt zu sprechen. «Menschen lernen ganz anders als Maschinen», sagt er zum Auftakt des Gesprächs per Video aus dem Homeoffice an der US-Westküste. Gemessen an seinem Bekanntheitsgrad ist er mit Mitte dreissig jung, und er wirkt gar noch etwas jünger. Der Experte für maschinelles Lernen arbeitet als Softwareentwickler bei Google und ist in der KI-Szene unter anderem bekannt, weil er ein verbreitetes Lehrbuch über maschinelles Lernen geschrieben hat. Vor allem aber, weil er vor vier Jahren diesen ARC-Datensatz entwickelt hat, an dem Entwicklerinnen und Forschende seither ihre Systeme messen – und sich die Zähne ausbeissen. Weil Menschen anders lernen als Maschinen.

Addition beispielsweise: Grosse Sprachmodelle müssen Millionen von Beispielen sehen, um überhaupt grob in der Lage zu sein, Zahlen zusammenzuzählen – und sie machen dennoch Fehler. «Einem Sechsjährigen kannst du Addition mit ganz wenigen Beispielen beibringen», sagt Chollet.

Ähnlich ist es mit den ARC-Aufgaben: Menschen genügen wenige Beispiele, um das jeweilige Konzept zu verstehen und es fortzuführen. Das liegt an unserer Lebenserfahrung, die sich in unsere Intuition eingräbt. So erkennen wir meist auf den ersten Blick das Konzept der jeweiligen Aufgabe: Hier wird ein Zwischenraum ausgefüllt, dort ein Bild gespiegelt, hier eine Bewegung fortgeführt, dort werden zwei Farben getauscht. Zentral sind die Fähigkeiten, zu abstrahieren und Schlussfolgerungen zu ziehen. Auch wenn die Aufgaben schwieriger werden, ist mit menschlicher Intuition leicht zu entscheiden, in welche Richtung es geht.

Chollets Ziel: Maschinen von ihrem Hang zur Inselbegabung zu befreien. Denn während wir Menschen breite, generelle Konzepte lernen – und das gewissermassen automatisch und von Geburt an durch unsere Interaktion mit der Welt –, werden Systeme künstlicher Intelligenz stets für eine konkrete Aufgabe entwickelt. Ihnen fehlt die Fähigkeit, zu abstrahieren und Schlüsse zu ziehen, um noch unbekannte Probleme lösen zu können. «Es zieht sich durch die Geschichte von KI», sagt Chollet seufzend: «Wir entwickeln Systeme für jeweils eine konkrete Aufgabe – und dann können sie auch nur diese eine Aufgabe erledigen. Das ist keine Intelligenz.»



Auch für sehr potente KI-Systeme ist diese Aufgabe nicht lösbar. Und für Sie? Hier können Sie weitere Testaufgaben lösen. Mensch schlägt Maschine. Und es macht Spass.

Aus seiner Sicht enden alle bisherigen Wege zur Entwicklung von KI in einer Sackgasse. Selbst die jüngsten Erfolge grosser Sprachmodelle können ihn nicht überzeugen. «Sprache enthält viel Wissen und Repräsentationen über die Welt», sagt er. Doch dieses Wissen reiche nicht, um die dahinterliegenden Konzepte zu verstehen. «Wir haben diesen Ansatz bis an seine Grenzen ausgereizt», sagt er. Als Trainingsdaten dient das ganze Internet, die neuesten Modelle sind riesig. «Es ist, wie einen Schwamm mit allem zu trainieren», sagt er. So beeindruckend die Ergebnisse seien, «es ist eine Sackgasse».

Kürzlich hat Chollet Chat GPT ebenso wie Googles Sprachmodell Bard eine Aufgabe vorgelegt, die ein ähnliches Muster hat wie jene aus seiner ARC-Challenge. Chat GPT gab eine Antwort, die offensichtlich lächerlich falsch war. Bard antwortete, es könne solche Aufgaben nicht lösen. «Bard gewinnt für Ehrlichkeit», tweetete Chollet.

In den vier Jahren, seit es die Challenge gibt, hätten sich die besten KI-Entwicklungsteams der Welt beteiligt, sagt François Chollet: «Der Test zieht die Leute an, weil er schwer ist und weil ihn noch niemand gelöst hat.»

Doch vielleicht braucht es gerade kein Elite-KI-Team – sondern eine ganz neue Herangehensweise. Das ist der Ansatz, den das Lab 42 in Davos verfolgt. «Ich glaube, es genügen zwei oder drei junge Menschen, Aussenseiter mit ungewöhnlichen Ideen», sagt Pascal Kaufmann: «Ich glaube, dass der Durchbruch in *human level AI* von einem Michelangelo, einem Einstein, einem da Vinci kommt.» Immer wieder, wenn Kaufmann von einem Raum in den anderen läuft, schaut er Michael Hodel kurz über die Schulter.

«Daten und Intelligenz sind Gegenpole»

Gerade arbeitet Hodel an einer Art Toolbox, einem Set von Funktionen, die Konzepte hinter einzelnen Aufgaben ausdrücken – zum Beispiel Objekte zu identifizieren oder eine Farbe gegen eine andere zu tauschen. Diese könne man zusammenschalten, um ARC-Aufgaben zu lösen, erklärt er. Allerdings würden das lange, aufwendige Konstruktionen. In einer schier endlosen Fleissarbeit hat er für jede der 400 Aufgaben ein eigenes Computerprogramm geschrieben: «Wenn ich unendlich viel Speicherkapazität und Laufzeit hätte, könnte ich jetzt fast alle Tasks lösen.» Der Wettbewerb verlangt jedoch nach schlanken Programmen, die in wenigen Stunden Lösungen bringen.

Eine Hoffnung: dass sein System aus den unzähligen Funktionen ein gewisses Verständnis für die Konzepte entwickeln kann – dass es also ähnlich wie Menschen relativ intuitiv erkennt, welches Konzept es nutzen muss, um eine Aufgabe anzugehen. Wird beispielsweise eine Fläche gefüllt, wird ein Objekt gespiegelt, bewegen sich Objekte aufeinander zu? Auch über diesen Umweg könnte doch noch maschinelles Lernen ins Spiel kommen, so sein Plan. Hodels andere Hoffnung: Mithilfe der Toolbox selbst unzählige weitere Trainingsdaten zu erzeugen, aus denen das System lernen kann.

«Aber Daten und Intelligenz sind ein Gegenpol», sagt Kaufmann zu Hodel, «wenn du viele Daten brauchst, dann hast du weniger Intelligenz. Oder willst du die Challenge mit *brute force* lösen?», fragt Kaufmann. Mit roher Gewalt also? Diesen Ausdruck nutzen Informatiker, wenn sie ausdrücken wollen, dass etwas durch viel Rechenaufwand oder viele Daten gelöst wird und eben nicht auf direkte, elegante Weise, mit so wenigen Daten wie möglich – also so, wie wir Menschen es machen.

Aber stimme die Aussage überhaupt, dass wir Menschen auf der Basis von wenigen Daten lernen, fragt Hodel zurück: «Ein Erwachsener hat ganz viel Vorwissen.» Nur deshalb erkennen wir die Konzepte hinter den Tasks auf den ersten Blick: weil wir seit unserer Geburt Daten verarbeiten und diese Konzepte aus dem Alltag kennen. Dieses ganze vorherige Wissen müsste aus Hodels Sicht fairerweise übersetzt werden in Trainingsdaten für KI, damit diese die gleiche Chance hat – so viel zur «schlanken Lösung».

Hodel ist mit seiner Idee nicht allein. Eine erfahrene Informatikerin aus den USA verfolgt einen ähnlichen Ansatz: Melanie Mitchell. Auch sie gilt als eine der Grossen in der Debatte um KI und Verstehen. Kürzlich hat sie ein Symposium ausgerichtet, in dem Informatikerinnen, Philosophen, Hirnforscherinnen und Fachleute für maschinelles Lernen mit Rang und Namen drei Tage lang über die Frage des Verstehens und KI diskutiert haben.

Was also ist menschliche Intelligenz?

Im Videointerview vor einem übervollen Bücherregal in ihrem Büro erklärt Melanie Mitchell – braune, halblange Haare, Brille, konzentrierter Blick –, dass sich das menschliche Wissen vermutlich nicht nur auf das beschränkt, was wir in unserem Leben bis dahin gesehen haben, sondern auch auf Erfahrungen vorheriger Generationen beruht, die in unseren Genen gespeichert sind. Und ausserdem auf sogenanntem kulturellem Wissen, das sich in menschlichen Kulturen bildet und verfestigt: «Eine Theorie ist, dass Menschen angeborene Konzepte haben, sogenannte *core concepts*», sagt sie. Diese Kernkonzepte helfen uns zu lernen, sie sind gewissermassen eine Abkürzung für uns.

Ähnlich wie Michael Hodel hat auch Mitchell neue ARC-Aufgaben erstellt, wenn auch deutlich einfachere, und diese nach Konzepten sortiert. Also beispielsweise «eine Fläche ausfüllen» oder «ein Objekt spiegeln». Ohne diese Kernkonzepte kommt KI aus ihrer Sicht nie auf ein menschliches Intelligenzniveau. Letztlich soll das dem Weg nahekommen, wie wir Menschen Konzepte lernen – nur dass wir es einfacher haben, weil wir einen Körper haben und Experimente machen können: «Babys probieren das einfach aus», sagt sie. «Man könnte Roboter bauen, die die kindliche Entwicklung nachempfinden.»

Diese Art maschinellen Lernens ist nicht neu, aber derzeit nicht besonders populär angesichts des Hypes um grosse Sprachmodelle. Auch, weil nach

wie vor viele überzeugt sind, dass diese kurz vor dem Durchbruch zu allgemeiner Intelligenz stehen.

Vielleicht können diese Kernkonzepte aber auch in Regeln übersetzt und einprogrammiert werden. Das allerdings sei nicht einfach, das zeige die Vergangenheit: «Wenn man zu viele Regeln vorgibt, ist das Programm nicht flexibel genug.» Bevor Computer genügend Leistungsfähigkeit hatten für maschinelles Lernen, hatten Programme vor allem durch Regeln funktioniert. Programmiererinnen mussten dafür alle Eventualitäten eingeben, die sich im Laufe der Zeit ergeben könnten. Im Sprachbereich hat sich aber schnell gezeigt, dass es kaum möglich ist, alles aufzuschreiben und in Regeln zu giessen, was die Welt zusammenhält – auch wenn Computerlinguisten über viele Jahre in langen Listen alles zusammentrugen, was Weltwissen ausmacht (beispielsweise: Jeden Menschen gibt es nur einmal; Vögel können fliegen und so weiter). Doch sie wurden nie fertig: Weltwissen ist schier endlos.

Als maschinelles Lernen in grossem Umfang technisch machbar wurde, verschob sich die Aufmerksamkeit: Seither verfolgen die meisten KI-Forscherinnen das Konzept, dass Maschinen aus Trainingsdaten Zusammenhänge selbst erkennen sollen. So funktionieren auch die grossen Sprachmodelle: Sie wurden mit vielen Terabyte an Daten aus dem Internet gefüttert – im Prinzip mit dem ganzen Internet – und lernten auf statistische Weise die Muster, die sich hinter unserer Sprache verbergen. Wenn sie nun als Chatbots eingesetzt werden, läuft im Hintergrund ein Prozess, der stets das wahrscheinlichste nächste Wort vorhersagt – so bilden sie Sprache.

Wir sehen Objekte, Computer nur Pixel

Ist das besser als die alten, regelbasierten Programme? «Maschinelles Lernen funktioniert recht gut», sagt Mitchell, «aber manche Dinge fehlen, die wir Menschen können, zum Beispiel abstrahieren.» Also eine Lösung auf ein anderes Problem übertragen und anpassen beispielsweise. Dabei helfen uns die angeborenen Kernkonzepte: Sie kommen uns so normal vor, dass es zunächst gar nicht so intuitiv ist, dass die Welt für Maschinen womöglich ganz anders aussieht: «Wenn wir unsere Umgebung betrachten, sehen wir Objekte, während Maschinen nur Pixel sehen», sagt Mitchell. Das Konzept dahinter heisst Segmentierung: Für uns ist klar, wo ein Objekt anfängt und wo es aufhört, «sogar Neugeborene erkennen das mühelos». Vermutlich können wir das, weil es wichtig ist für unser Überleben: «Es ist nützlich, die Welt in Objekte aufteilen zu können.»

So erkennen wir auch, wenn ein Objekt hinter einem anderen verschwindet, weil es sich dorthin bewegt – während die Veränderungen durch Bewegungen von Objekten für ein Computerprogramm lediglich aussehen, als ob sich Pixel verschieben und diese schliesslich teilweise verschwinden oder die Farbe verändern. «Vieles von diesem tiefen Wissen über die Welt ist nicht einmal aufgeschrieben», sagt die Informatikerin.

Das heisst, grosse Sprachmodelle können sich zwar auf Basis von Sprachmustern vieles über die Welt «zusammenreimen», aber entscheidendes Wissen ist so tief in uns Menschen verankert, dass wir es nie in Worte fassen. Dafür bleiben Sprachmodelle blind. «Deshalb versagen sie in vielen Aufgaben des logischen Denkens und des Generalisierens», sagt Mitchell. Und deshalb werden Sprachmodelle wie Chat GPT aus ihrer Sicht nie den Weg zu allgemeiner künstlicher Intelligenz ebnen. Sie bräuchten ein Verständnis für die Konzepte, die in der ARC-Challenge geprüft werden – doch das haben sie nicht.

Ob die Lösung der ARC-Challenge hinreichend ist für menschenähnliche Intelligenz, steht allerdings ebenfalls infrage. «Wenn Computer tatsächlich in der Lage sein sollten, es zu erkennen, wenn sich beispielsweise ein Objekt über das andere bewegt – verstehen sie dann wirklich, was das bedeutet?», fragt Mitchell. Nur weil das ein Programm in einem Fall richtig mache, sei noch lange nicht erwiesen, dass es das dahinterliegende Konzept verstehe – was wiederum die Grundlage dafür wäre, dass es zuverlässig funktioniert.

Mitchell hat das kürzlich demonstriert. Sie war genervt, weil in den sozialen Netzwerken immer wieder Beispiele kursierten, die angeblich zeigten, dass die grossen Sprachmodelle die Welt tatsächlich «verstehen», dass sie physikalische Zusammenhänge richtig einordnen können und vieles mehr. «Das sind immer Einzelbeispiele. Sobald man das anders formuliert, merkt man, dass das Modell nichts verstanden hat.» In ihrem Beispiel schlug das Modell vor, eine Zahnbürste in Pudding zu stecken, darauf ein Marshmallow zu balancieren und auf diesem ein Glas Wasser. Jeder Mensch kann sich vorstellen, dass das nicht funktionieren kann.

Aber was ist denn überhaupt «verstehen»? Wie machen wir Menschen das?

Oder sind wir auch einfach bloss Vorhersage-Maschinen?

«Being you» – der Schriftzug prangt über einer riesigen Darstellung der Iris des menschlichen Auges. Es ist das Cover des Buches von Anil Seth: auf mehr als 350 Seiten hat der Professor für Cognitive and Computational Neuroscience an der University of Sussex ausgeführt, wie das menschliche Bewusstsein funktioniert. «Noch ist es relativ einfach, Sprachmodelle bei Fehlern zu erwischen, die zeigen, dass sie nicht wirklich verstehen», sagt Seth, kahlrasiert und in schwarzem T-Shirt, im Videocall. Das werde sich aber in Zukunft ändern, wenn diese Modelle immer besser würden. Letztlich seien sie aber einfach nur *«next-token prediction machines»,* sagt er. Also Modelle, die vorhersagen, welches Wort auf das nächste folgt – allein auf Basis von statistischer Wahrscheinlichkeit.

«Man könnte provokativ behaupten, dass das alles ist, was Verstehen ausmacht, und dass das in unserem Kopf genauso funktioniert», sagt Seth. Er selbst gehe zwar nicht davon aus, dass es so sei – «aber es ist auch immer eine Überlegung wert». Denn: Auch unser Gehirn ist eine Vorhersagemaschine.

Immer wieder hat die Hirnforschung gezeigt, dass wir permanent im Alltag vorhersagen, was geschieht. Schalten wir das Licht an, wird es heller im Raum. Lassen wir etwas fallen, schlägt es auf dem Boden auf. Unser Gehirn entwirft ständig Hypothesen über die Welt und überprüft sie. Nur wenn sie falsch sind, wenn etwas anderes passiert als erwartet, passen wir das Modell an. Funktionieren menschliche Gehirne und grosse Sprachmodelle mit ihrem Vorhersageprinzip also doch ähnlicher, als wir uns eingestehen wollen?

«Es gibt viele Parallelen zu künstlicher Intelligenz», sagt Anil Seth, «aber die Versuchung ist gross, uns für etwas Besonderes zu halten.» Noch ist aus neurowissenschaftlicher Sicht völlig unklar, was Verstehen genau bedeutet oder wie das im biologischen Gehirn vor sich geht. Vielleicht ist es doch nicht mehr als die Vorhersagemaschine, die wir in künstlichen neuronalen Netzen beobachten können.

Eines, so Seth, sei sicher: «Wenn du ein Konzept verstanden hast, dann kannst du es auch generalisieren.» Sprich: Wenn wir sehen, dass das Konzept darin besteht, dass eine Fläche ausgefüllt wird wie in den ARC-Aufgaben, können wir das auch dann umsetzen, wenn die nächste Fläche grösser ist oder eine andere Form hat. «Ein Psychologe würde aber sagen, dass wir selbst etwas, das wir sehr gut generalisieren können, noch nicht unbedingt verstehen.» Denn eine weitere Grundlage für Verstehen sei die körperliche Dimension. Seth greift nach seiner Kaffeetasse, die vor ihm auf dem Tisch steht: «Ich kann diese Tasse hochnehmen und verstehe, was das bedeutet.» Zu spüren, wie sie sich anfühlt, ihre Form, ihr Gewicht – all das hilft auch beim Verstehen.

Sprachmodelle haben keinen Körper. Und in der Tat zeigen sich Chatbots immer wieder «verwirrt», wenn es um physikalische Zusammenhänge geht – wie ja auch Mitchells oben beschriebenes Experiment mit der Zahnbürste und dem Marshmallow zeigt. Auch in diesem Fall ist es für Menschen auf den ersten Blick klar, wieso das nicht funktionieren kann – wahrscheinlich auch, weil wir unzählige haptische Erfahrungen mit Gegenständen und Begegnungen mit der Schwerkraft hinter uns haben. «Deshalb fühle ich mich der konservativen Herangehensweise näher, die besagt, dass grosse Sprachmodelle das Verstehen nur simulieren», sagt Seth.

Für echtes Verstehen brauche es vermutlich einen Körper. «Wir sind evolutionär vortrainiert», sagt er und nutzt wieder ein Wort, das auch im Zusammenhang mit neuronalen Netzen verbreitet ist – auch diese sind vortrainiert und werden dann auf einen speziellen Anwendungsfall feintrainiert. Es steckt sogar im Namen Chat GPT, *General Pretrained Transformer*. Ein Unterschied dabei sei, sagt Seth: «Unser Vortraining fand auf verkörperte Art und Weise statt.»

Dennoch könne es sein, dass Maschinen anders verstehen. Wir sollten nicht zu sehr von uns ausgehen, wenn wir Allgemeine Künstliche Intelligenz bauen wollen, warnt Seth zum Abschluss des Gesprächs: «Wir sollten uns immer im Klaren sein, dass es darum geht, ein Werkzeug zu bauen, nicht einen ‹Kollegen›.» Von daher könnte das Abstrahieren in Maschinen auch anders funktionieren – nicht alles muss an den Menschen angelehnt sein.

Die Suche nach neuen Wegen

Am zweiten Tag meines Besuchs in Davos sitzt Michael Hodel nachdenklich vor seinen Bildschirmen in der Lab-42-Villa. Er hat sich gerade mit dem aktuellen Gewinner der ARC-Challenge per Linkedin ausgetauscht. «Bist du weiter dran?», wird er von ihm gefragt. Hodel seufzt vor seinem Bildschirm. Er kann nicht anders. «Ich bin angefressen von der Challenge.» Aber ist die Lösung der ARC-Challenge überhaupt der richtige Weg, um zu menschenähnlicherer KI zu kommen?

Bei seiner Internetsuche nach Antworten landete der Student schliesslich bei einer Erklärung des Gewinners der ARC-Challenge von vor drei Jahren. «Icecuber» lautet sein Alias auf der Plattform Kaggle. Dahinter verberge sich ein schlauer norwegischer Jugendlicher, sagt Hodel und klingt bewundernd. Icecuber erklärt seine Lösung – sie hat Gemeinsamkeiten mit dem Ansatz Hodels – und schreibt darunter: «Leider habe ich nicht das Gefühl, dass meine Lösung uns der Allgemeinen Künstlichen Intelligenz näherbringt.» Unter seinem Beitrag diskutieren andere User über seinen Weg und der Challenge. Einer schreibt tröstend: «Ich glaube, wenn du so weiterarbeitest, kannst du ARC vielleicht auch ohne Allgemeine KI knacken.»

Was, wenn die Kernkonzepte doch nicht wichtig sind für Allgemeine-KI? Wenn die Lösung der Challenge also tatsächlich nicht zu menschenähnlicher maschineller Intelligenz führt? «Das ist mir egal», sagt Hodel fast trotzig. «Mir geht es nicht um Allgemeine KI, mir geht es um die Challenge.»

Pascal Kaufmann hingegen jagt den Braincode, für ihn ist die Challenge nur Mittel zum Zweck. Deshalb bespricht er jetzt im Boardroom mit seinem Team die nächsten Schritte für die Entwicklung von menschenähnlicher KI. Zur Wahl stehen: eine KI, mit der ein Avatar in einer virtuellen Welt ausgestattet ist – oder Schwarmintelligenz. Der Avatar soll der KI einen Körper geben in der Hoffnung, dass das der fehlende Baustein ist. «Es muss ein kleiner neugieriger Wissenschaftler sein, der sich sein Wissen selbst aneignet», sagt Kaufmann am grossen runden Tisch. Vielleicht könnte die KI auf diese Weise auch Angst lernen, denn ein Körper ist verletzlich. «Emotionen sind eine Brücke», sagt Kaufmann. Denn Emotionen bündeln Dinge. «Anstatt 300 Millionen Sensoren abzurufen, sagt mein Gefühl: Geh da besser weg.»

Für die andere Hoffnung – Schwarmintelligenz – konzipiert das Team gerade einen neuen Wettbewerb. «Schwärme haben Eigenschaften, die ich in neuronalen Netzen noch nicht sehe», sagt Kaufmann. Der Mensch bestehe schliesslich auch aus vielen Zellen, die sich nach noch unbekannten Regeln selbst organisierten. «Wir sind ein Superorganismus.» Mit einem entsprechenden Wettbewerb, ist Kaufmann überzeugt, gelinge es dem Lab bestimmt, «das Prinzip der Intelligenz zu finden». Ein Mitarbeiter berichtet, dass er alle wichtigen Schwarmintelligenz-Forscherinnen angeschrieben habe, viele davon Robotiker, «sie haben grosses Interesse, mit uns zu reden».

Wer zwei Tage im Lab 42 verbringt, bekommt den Eindruck, dass die Suche nach menschenähnlicher Intelligenz einem Stochern im Nebel gleicht. Vielleicht weil wir gar nicht genau wissen, wonach wir suchen. «Das ganze Konzept um Allgemeine KI ist ohnehin so unklar», sagt Melanie Mitchell. Sogar in Bezug auf uns Menschen selbst: «Die meisten Kognitionswissenschaftler würden sagen, dass Menschen gar keine allgemeine Intelligenz haben. Zum Beispiel, weil wir schlecht darin sind, mit Wahrscheinlichkeiten umzugehen.»

Wir seien aber auch noch in etwas anderem schlecht, sagt Neurowissenschaftler Anil Seth: darin, KI zu verstehen. Die neuen grossen Sprachmodelle verhalten sich nämlich sonderbar, sie sind für Menschen überhaupt nicht vorhersehbar: «Manche werden nach vielem Trainieren wieder schlechter, manche generalisieren besser als andere – aber wir verstehen überhaupt nicht, warum das so ist.» Selbst Fachleute sind verblüfft, welche Massnahmen die Modelle besser und welche sie schlechter werden lassen.

Vielleicht führt der Weg zu menschenähnlicher KI also erst mal darüber: dass Menschen die Maschinen besser verstehen, die sie gebaut haben.

Zur Autorin

Eva Wolfangel ist freie Journalistin und schreibt über künstliche Intelligenz, virtuelle Realität und Cybersecurity. Eine ihrer Leitfragen lautet: Wie leben wir in Zukunft? Sie wurde 2018 als «European Science Writer of the Year» ausgezeichnet.