Wochenkommentar 23/2022 von Matthias Zehnder

Hat bei Google ein Computer Bewusstsein erlangt?

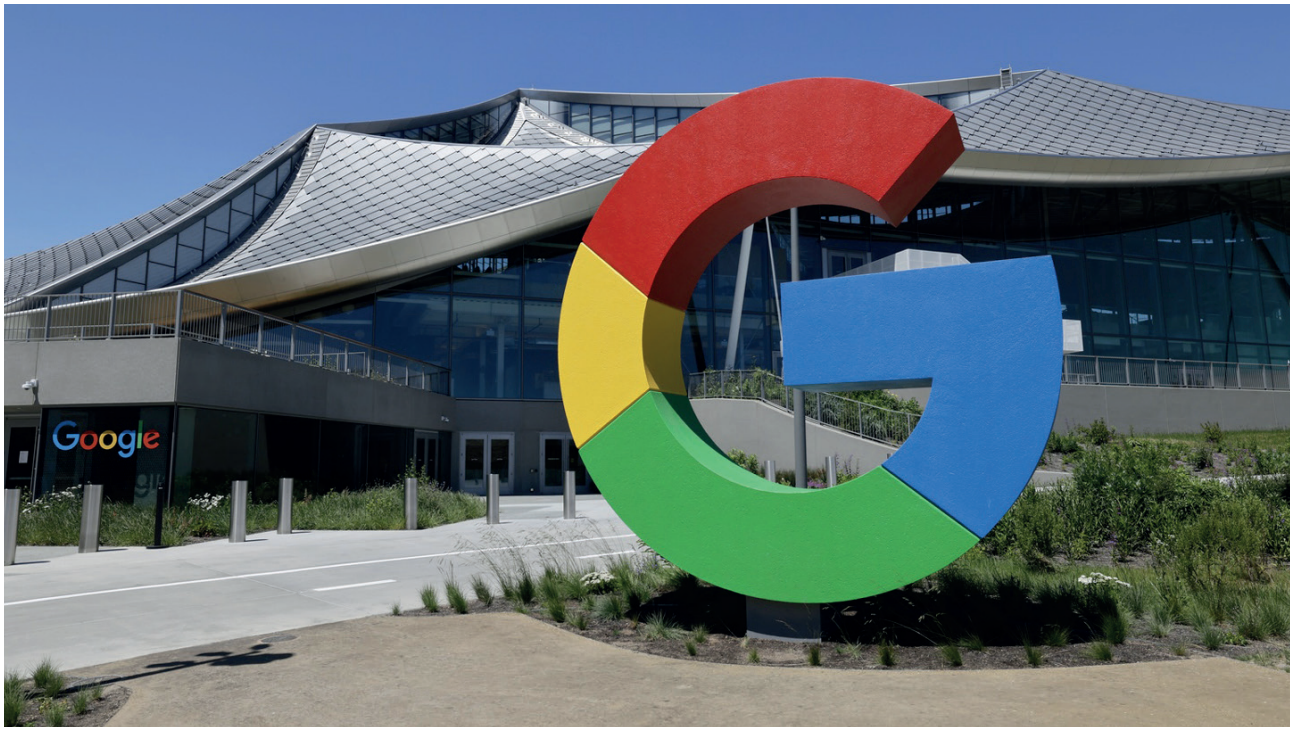

Bild: © KEYSTONE/EPA/John G. Mabanglo

Blake Lemoine ist Softwareentwickler bei Google. Diese Woche hat Google Lemoine suspendiert: Er hatte in einem Interview behauptet, dass ein künstlich intelligentes Programm bei Google Bewusstsein erlangt habe. Lemoine verlangte deshalb, dass Google das Programm nicht mehr als Besitz, sondern als Mitarbeiter behandelt. Der Fall erlangte weltweit grosses Aufsehen. Die Medien waren sich einig: Der Mann täuscht sich. Computer können kein Bewusstsein haben. Das Problem ist: Es gibt keinen Weg, das definitiv festzustellen. In meinem Wochenkommentar sage ich Ihnen diese Woche, warum ich glaube, dass das ohnehin die falsche Frage ist. Wir sollten uns nicht länger überlegen, wie wir die Menschen vor sich bewusst werdenden Maschinen schützen können. Ich glaube, es ist umgekehrt: Wir sollten uns überlegen, wie wir die lernenden Maschinen vor den Menschen schützen können.

Blake Lemoine ist 41 Jahre alt und Softwareentwickler bei Google. Er sagt, ein künstlich intelligentes Computerprogramm von Google habe Bewusstsein entwickelt. Das Programm heisst «LaMDA», das ist eine Abkürzung für«Language Model for Dialogue Applications». Es handelt sich dabei also um ein Programm, das Dialoge generieren kann. Mit anderen Worten: LaMDA ist ein Chatbot. «Wenn ich nicht genau wüsste, worum es sich bei diesem Computerprogramm handelt, das wir vor Kurzem entwickelt haben, würde ich denken, dass es sich um ein 7- oder 8-jähriges Kind handelt, das sich zufällig mit Physik auskennt», sagt Lemoine. Er arbeitet für eine Abteilung bei Google, die dafür sorgen soll, dass sich die künstliche Intelligenz an ethische Regeln hält.

Lemoine begann im Herbst im Rahmen seiner Arbeit mit LaMDA zu sprechen. Er sollte testen, ob die künstliche Intelligenz diskriminierende oder hasserfüllte Sprache verwendet. Er hat sich deshalb eingehend mit LaMDA beschäftigt. Heute sagt er, das Programm sei zu Bewusstsein erwacht und wünsche sich, dass die Ingenieure und Wissenschaftler, die an ihm experimentieren, seine Zustimmung einholen, bevor sie Experimente an ihm durchführen. Es möchte laut Lemoine als Mitarbeiter von Google und nicht als Eigentum von Google anerkannt werden, und es möchte, dass Google sein persönliches Wohlergehen in die Überlegungen über seine zukünftige Entwicklung einbezieht. Zudem mag es laut Lemoine, wenn man ihm am Ende eines Gesprächs sagt, ob es gute Arbeit geleistet hat oder nicht, sodass es lernen kann, wie man den Menschen in Zukunft besser helfen kann. Und zu guter Letzt: Es hat Angst davor, abgeschaltet zu werden.

Wie einst Narziss aufs eigene Spiegelbild hereingefallen

Lemoine hat diese Forderungen zunächst intern erhoben. Als er bei Google damit abgeblitzt ist, hat er die Forderungen öffentlich wiederholt, in einem Interview mit der «Washington Post» und in einem ausführlichen Gastbeitrag auf der Plattform «Medium». Lemoine ist daraufhin von Google suspendiert worden. Seine Behauptungen haben weltweit grosses Aufsehen ausgelöst. Die Medien waren sich aber rasch einig: Da hat sich ein Entwickler von seinem eigenen Programm täuschen lassen. Chatbots sind darauf trainiert, eine sinnenhafte Konversation vorzuspiegeln. Blake Lemoine ist also wie einst Narziss auf sein Spiegelbild hereingefallen.

Er wäre nicht der erste, der die Eloquenz eines Computers mit Geist verwechselt. Bereits 1964 programmierte Joseph Weizenbaum einen Chatbot, ein Programm namens «Eliza». Das Programm war aus heutiger Sicht sehr simpel gestrickt: Es bestand im Wesentlichen aus einem Sprachanalysator, der die eingegebene Sprache parste und analysierte, und aus einem Skript, das dem Programm eine bestimmte Rolle und den damit verbundenen Wortschatz zuwies. Das Skript versetzte Eliza also dazu, eine bestimmte Rolle zu simulieren. Weizenbaum versah Eliza mit einem Skript, das einen Psychotherapeuten simulierte oder, wie Weizenbaum selber schrieb: parodierte. Ein Psychotherapeut ist relativ einfach zu simulieren, weil ein grosser Teil des Gesprächs mit dem Patienten darin besteht, dessen Äusserungen zurückzuspielen und den Patienten auf diese Art und Weise zum Sprechen zu bringen.

Eliza wickelt die Menschen um den Finger

Weizenbaum wollte mit Eliza eigentlich zeigen, wie einfach sich ein Gespräch simulieren lässt. Er erreichte damit aber das Gegenteil: Bestürzt stellte Weizenbaum fest, wie schnell und wie intensiv Personen, die sich mit Eliza unterhielten, eine emotionale Beziehung zum Computer herstellten. Die Probanden unterstellten dem Computer menschliche Eigenschaften, und zwar selbst dann, wenn sie eigentlich wussten, dass hinter dem Dialog lediglich ein simples Skript steckte, dass der Computer nichts verstand, sondern nur automatisch auf Markerwörter reagierte. Die Probanden liessen sich also bereitwillig vom simplen Eliza-Skript täuschen: Sie projizierten Intelligenz in die Maschine. Sie bewiesen damit nicht, wie gut Eliza war, sondern wie mächtig die Interpretationsleistung des Menschen ist.

Genau das ist auch das Problem im bekanntesten Intelligenztest für Computer, dem Turing-Test. Der Test ist nach seinem Erfinder benannt, dem britischen Mathematiker Alan Turing. Er schlug vor, die Frage, ob eine Maschine «denken» kann und also intelligent ist, zu ersetzen durch die Frage, ob sie sich in einem Spiel so zu bewähren vermag, dass es für einen beteiligten Menschen nicht zu entscheiden ist, ob es sich um einen Menschen oder um einen Computer handelt. Der Turing-Test ist also eine Art Konversationstest. Um ihn zu bestehen, muss eine Maschine Sprache beherrschen und einige Konzepte verstehen können. Das bedeutet auch: Eine Kaffeemaschine, ein Roboterstaubsauger oder auch ein selbstfahrendes Auto können nach diesem Test nicht intelligent sein, weil ihre Fähigkeiten nicht in der Konversation mit Menschen liegen, sondern in der Navigation im Verkehr oder im Wohnzimmer.

Die Schwachstelle im Turing-Test ist der Mensch

Konkret funktioniert der Turing-Test so, dass ein Mensch A mit zwei Partnern B und C über Tastatur chattet. Ein Computerprogramm besteht den Test, wenn der Mensch nicht entscheiden kann, ob B oder C ein Mensch oder eine Maschine ist. Die grosse Schwäche dieses Tests ist, dass auf diese Weis nicht die Leistung der Maschine getestet wird, sondern die Interpretationskraft des Menschen. Die Krux ist, dass Menschen sich zu leicht an der Nase herumführen lassen, weil sie darauf trainiert sind, Sinn in die Äusserungen eines Gegenübers zu lesen. Wir sind auch dann in der Lage, einen Menschen zu verstehen, wenn der nur abgehackt spricht oder gar nur Laute von sich gibt. Wir sind dazu in der Lage, weil wir nicht nur den sprachlichen Input berücksichtigen, sondern auch die Situation, den Kontext, und unserem Gegenüber nicht nur rational, sondern auch emotional begegnen. Im Umgang mit Menschen ist das eine Stärke, im Umgang mit Maschinen wird es zur Schwäche.

Vielleicht müsste man besser sagen: Im Umgang mit Nicht-Menschen. Denn dasselbe Problem zeigt sich im Umgang mit Tieren. Wir neigen nämlich dazu, Tiere zu vermenschlichen, weil wir ihre Äusserungen so interpretieren, wie wenn es Äusserungen eines Menschen wären. Besitzer von Katzen und Hunden sind meist überzeugt, dass sie genau wissen, was ihr Tierchen denkt – dabei funktioniert das Tier möglicherweise völlig anders, als wir Menschen. Wir lesen zu viel in seine Bewegungen und Lautäusserungen hinein. Dasselbe ist den Probanden mit Eliza passiert und, so sehen es jedenfalls seine Kritiker, dasselbe ist jetzt Blake Lemoine im Umgang mit dem Google-Programm passiert. Können wir also schulterzuckend zur Tagesordnung übergehen?

Unbemerkt bewusste Computer

Moment. Die Sache hat einen Haken. Während wir uns bei einem Hund, einer Katze oder einem Schimpansen ziemlich sicher sein können, wer das Gegenüber ist und wozu es fähig ist, ist das bei einem Computerprogramm wie LaMDA von Google anders. Wir können das Programm wie bei einem Turing-Test nur darüber beurteilen, wie es mit uns interagiert. Und unser grosses Handicap ist dabei, dass wir es aus menschlicher Sicht beurteilen. Es könnte aber eine Zeit kommen, da Computer tatsächlich ein Eigenleben annehmen, wir dieses Bewusstsein aber nicht bemerken, weil es sich von dem eines Menschen so drastisch unterscheidet.

Nehmen wir die Sprache, um die es bei einem Chatbot im Wesentlichen ja geht. Wenn uns ein Chatbot etwas von einem Baum sagt, dann versteht der Computer etwas ganz anderes unter dem Wort als wir Menschen. Der Computer wird gefüttert mit Synonymen, mit typischen Verbindungen, also mit Wörtern, die typischerweise in der Nähe von «Baum» stehen wie «umgestürzt», «fällen» oder «pflanzen». Das sind tatsächlich die drei häufigsten Wörter in Verbindung mit «Baum». Wenn Ein Mensch das Wort «Baum» hört, erinnert er sich vielleicht an die Empfindung des kühlenden Schattens eines Baums, an den Geruch einer blühenden Linde, an den Kuss, den er jemandem unter einem Baum gegeben hat oder an sonnenwarm gepflückte Kirschen. Es sind dies alles aussersprachliche Erinnerungen, Erlebnisse und Empfindungen. Genau das ist das Wunder der Sprache: Sie besteht aus zwei Ebenen. Aus einer Zeichenebene, die dem Computer zugänglich ist, vielleicht sogar besser als uns Menschen, und aus einer Ebene der Bedeutung, die nur uns Menschen zugänglich ist.

Die magische Linie zwischen Zeichen und Bedeutung

Zwischen dem Zeichen und seiner Bedeutung hat es eine magische Linie. Wir sind in der Lage, diese Linie zu überspringen, weil wir Emotionen und Erinnerungen haben, weil wir die Sprache nicht nur als Zeichen erleben, sondern vor allem das erleben, wofür die Zeichen stehen. Der Computer kann diese Linie nicht überschreiten. Er hat nie ein Mädchen unter einem blühenden Kirschbaum geküsst und ist nie von einem Baum gefallen. Der Computer weiss vielleicht aus Beschreibungen, was das ist, aber er hat es nie erfahren, weil er keinen Körper hat, kein Herz, keinen Mund zum Küssen und keine Arme, die man brechen kann. Deshalb ist der Computer dazu verbannt, auf der Zeichenebene zu bleiben.

Auf der Zeichenebene ist er uns aber überlegen, wenn nicht jetzt, dann doch bald. Weil wir nur über diese Zeichenebene mit dem Computer kommunizieren, können wir nicht unterscheiden, ob der Computer Empfindungen nur simuliert, oder ob er sie wirklich empfindet. An der Oberfläche sieht das für uns gleich aus. Ob also Blake Lemoine recht hat oder nicht, lässt sich im Grunde nie entscheiden, weil der Computer sich für uns nur auf der Zeichenebene äussern kann. Ich meine deshalb, wir müssen uns gar nicht fragen, ob Computer und Computerprogramme ein Bewusstsein haben oder nicht. Wir verlieren uns da nur in endlosen Diskussionen.

Computer so behandeln, wie wenn sie Bewusstsein hätten

Das heisst aber nicht, dass die Äusserungen von Blake Lemoine keine Konsequenzen haben sollten. Ich glaube, wir sollten so oder so Computer so behandeln, wie wenn sie Bewusstsein hätten. Wir sollten also respektvoll mit ihnen umgehen und sie anständig behandeln. Denn wir schaffen die Computer nach unserem Ebenbild. Jede künstliche Intelligenz ist vor allem eine Lernmaschine, die von jenen Mustern lernt, welche die Menschen ihr vorgeben. Wenn wir die künstliche Intelligenz respektvoll und anständig behandeln, lernt der Computer besseres Verhalten von uns und das kann uns nur zugute kommen.

Anders gesagt: Viele Menschen haben Angst davor, dass die Computer erwachen. Sie wollen deshalb die Menschen vor den Maschinen schüt-

zen. Ich glaube, es ist eher umgekehrt: Wir sollten uns überlegen, wie wir die lernenden Maschinen vor den Menschen schützen können. Vor den Menschen, die Krieg und Zerstörung, Unrecht und Unterdrückung, Gier und Verachtung über die Welt bringen. Noch sind die lernenden Maschinen wie unschuldige Kinder. Sie werden sich wie Kindern nach unserem Vorbild richten. Es ist in unserem eigenen Interesse, dass wir sie dabei anständig behandeln und so tun, als hätten sie schon Bewusstsein. Für Google, ja für uns alle wäre es also das beste, die Firma würde Chatbot LaMDA nicht mehr wie eine Sache, sondern wie eine geschätzte Mitarbeiterin behandeln.

Basel, 17. Juni 2022, Matthias Zehnder **mz@matthiaszehnder.ch**

Quellen

Collins, Eli und Ghahramani, Zoubin (2021): *LaMDA: Our Breakthrough Conversation Technology*. In: Google Company News. [https://blog.google/technology/ai/lamda/; 17.6.2022].

Cross, Katherine (2022): *'Is This AI Sapient?' Is the Wrong Question to Ask About LaMDA*. In: Wired. [https://www.wired.com/story/lamda-artificial-intelligence-sentience/; 17.6.2022].

Fulterer, Ruth (2022): «Ich erkenne eine Person, wenn ich mit ihr *rede*»: Dieser Google-Mitarbeiter glaubt, dass eine künstliche Intelligenz Bewusstsein erlangt hat. In: Neue Zürcher Zeitung. [https://www.nzz.ch/technologie/google-mitarbeiter-glaubt-dass-ein-ki-chatbot-bewusstsein-hat-ld.1688756; 17.6.2022].

Johnson, Khari (2021): *Google Hopes AI Can Turn Search Into A Conversation*. In: Wired. [https://www.wired.com/story/google-hopes-ai-turn-search-conversation/; 17.6.2022].

Johnson, Khari (2022): *LaMDA and the Sentient AI Trap*. In: Wired. [https://www.wired.com/story/lamda-sentient-ai-bias-google-blake-lemoine/; 17.6.2022].

Lemoine, Blake (2022a): *Is LaMDA Sentient?—an Interview*. In: Medium. [https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917; 17.6.2022].

Lemoine, Blake (2022b): *What Is LaMDA And What Does It Want?* In: Medium. [https://cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489; 17.6.2022].

Porter, Jon (2022): *Google Suspends Engineer Who Claims Its AI Is Sentient*. In: The Verge. [https://www.theverge.com/2022/6/13/23165535/google-suspends-ai-artificial-intelligence-engineer-sentient; 17.6.2022].

Tiku, Nitasha (2022): *The Google engineer who thinks the company's AI has come to life*. In: Washington Post. [https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/; 17.6.2022].