Illustration iStock

Es gibt kein Auslöschungsrisiko durch diese Künstliche Intelligenz

Sogar renommierte Fachleute warnen vor der KI – und vergleichen deren Risiken mit denen von Atomwaffen. Warum das zu weit geht.

Von Ralf Otte

Um das Risiko der gegenwärtig existierenden Künstlichen Intelligenz (KI) besser zu verstehen, bleibt nichts anderes übrig, als die aktuellen Diskussionen zu erweitern und ins technische Detail zu gehen. Daher zunächst drei technische Aussagen, die zentral sind: Erstens hat die heutige KI kein Bewusstsein. Zweitens hat sie keine Gefühle. Drittens hat sie keinen Willen.

Über alle drei Begriffe ist hinlänglich diskutiert worden, und tatsächlich hängt von ihrem Verständnis viel ab. Wer das Fundament der Künstlichen Intelligenz nicht versteht, wird in diesem Bereich letztlich vieles glauben (müssen), wenn es nur gut argumentiert ist.

Betrachten wir zuerst ein System mit Bewusstsein, zum Beispiel einen Menschen, der gerade eine Zeitung liest. Während des Lesevorganges fallen Lichtstrahlen vom Papier über die Augenlinse auf die Netzhaut. Am Ort des Auftreffens der elektromagnetischen Welle gibt es biochemische Veränderungen, wodurch elektrische Signale erzeugt werden, die entlang von Nervenbahnen ins Gehirn gleitet werden. Am Ende der Verarbeitungskette finden wir zahlreiche neuronale Netzwerke im visuellen Cortex, die nun erregt sind und ein (neuronales) Abbild der Zeitung im Gehirn codieren. Bei Systemen mit Bewusstsein passiert indes weit mehr. Jeder Mensch hat neben dem Abbild im Gehirngewebe noch ein weiteres Abbild – in seinem Bewusstsein, und dieses „Bewusstseinsabbild“ sorgt dafür, dass jeder Mensch die Zeitung vor seinem Kopf wahrnimmt und zwar exakt dort, wo sich die Zeitung auch befindet, nämlich etwa auf dem Tisch.

Dieses visuelle Hochleistungsphänomen, das wir als selbstverständlich hinnehmen, kann das Gehirngewebe alleine nicht leisten, denn es gibt keinen neuronalen Mechanismus, der das codierte Abbild aus dem visuellen Cortex des Gehirngewebes nach außen in die Umwelt projiziert. Die physikalische Erklärung dieses Phänomens, dass Menschen sozusagen aus ihren Augen hinaus schauen, ist nichttrivial. Es kann auch nicht mit Lernprozessen in der Kindheit erklärt werden, sondern basiert auf einem sonderbaren physikalischen Effekt: Das Nachaußenschauen wird mithilfe bestimmter physikalischer Aspekte des Bewusstseins realisiert. Ein normaler Computer kann aus seiner Kamera nicht in die Umwelt hinaussehen. Und zwar prinzipiell nicht. Er hat immer nur ein Zahlenabbild seiner Umgebung in seinem Speicher abgelegt, das anschließend als Bild interpretiert wird.

Gegenwärtige KI-Systeme verfügen über eine interne Repräsentation ihrer Umgebung, Systeme mit Bewusstsein, wie Menschen eben, sehen die Umwelt wirklich im Außen, dies nennt der Fachmann Wahrnehmung. Wenn Sie demnächst einen Freund ansehen, wundern Sie sich also nicht, dass das Bild Ihres Freundes vor Ihrem geistigen Auge direkt auf seinem Körper erscheint (und nicht in Ihrem visuellen Cortex im Hinterkopf). Eine solche Wahrnehmung ist für heutige Computersysteme etwas völlig Unmögliches und übrigens ein Hauptproblem des autonomen Fahrens. KI-Systeme können nicht wahrnehmen wie Menschen, sie simulieren das nur. Simulationen sind bis zu einem gewissen Grade auch völlig in Ordnung, aber ab einer gewissen Komplexität der realen Umgebung, in der die Simulationen nutzbringend angewendet werden sollen, scheitern sie zwangsläufig. Das ist keine Kleinigkeit.

Nun vom Sehen zur Sprache: KI-Systeme können Sprache nicht verstehen wie wir Menschen, sondern auch hier das Verständnis nur simulieren. Der amerikanische Philosoph John Searle hat das Problem des Verstehens mit seinem Gedankenexperiment vom sogenannten Chinesischen Zimmer bekanntlich schon hinreichend genau ausbuchstabiert. Man kann so tun, als ob man eine Sprache verstünde, wenn man auf ein Eingangsmuster das jeweils richtige Antwortmuster präsentiert, auch wenn man weder das Eingangsmuster (zum Beispiel ein chinesisches Schriftzeichen), noch das Ausgangsmuster (abermals ein chinesisches Schriftzeichen) auch nur im Ansatz versteht. Den Entwicklern der Sprachroboter ist es nun gelungen, für wichtige Sprachen der Welt die ihnen zugrundeliegenden mathematischen Beziehungen zu extrahieren. Und – besonders wichtig –, da sie dies mit künstlichen neuronalen Netzen umsetzen, müssen diese die Beziehungen nicht einmal explizit selber kennen, sondern können die „Mathematik der Sprache“ an Milliarden von Texten implizit erlernen. Einen ähnlichen Durchbruch gab es vor Jahren mit sogenannten neuronalen Faltungsnetzen (CNN) im Bereich des maschinellen Sehens. Über die Restriktionen dieses „mathematischen Sehens“ wurde oben schon gesprochen.

Der Chatbot ChatGPT hat Millionen Menschen rund um den Globus erstmals in Kontakt gebracht mit dem, was moderne KI-Systeme können. Viele sind überrascht, erschrocken, die Gesellschaft ist regelrecht erregt. So ausführlich und intensiv hat die breite Öffentlichkeit jedenfalls seit vielen Jahren nicht mehr über KI diskutiert. Mit manchen scheint indes die Phantasie durchgegangen sein, mitunter warnen Zeitgenossen schon vor einer möglichen Machtübernahme durch die KI. Sogar renommierte KI-Forscher und Unternehmer sorgen sich um die womöglich heraufziehenden Gefahren, darunter nicht zuletzt Sam Altman selbst, der Chef der Unternehmung Open AI, die ChatGPT entwickelt hat. Auch er meint, sie besitze „eine globale Priorität neben anderen gesellschaftlichen Risiken wie Pandemien und Atomkrieg“. Diese Warnung könnte von großer Sorge herrühren – allerdings auch auf der Hoffnung begründet sein, extrem viel Aufmerksamkeit zu erzielen und das eigene KI-System sehr mächtig erscheinen zu lassen, und vielleicht selbst Teil einer internationalen KI-Regulierungskommission zu werden, mit möglichen Nachteilen für (potentielle) Wettbewerber.

Doch solch extreme Angst zu verbreiten ist fehl am Platze. Seit Jahrzehnten wird eine KI in der Industrie eingesetzt, die genauso intelligent ist, wie die, die gerade ins Rampenlicht tritt. Und auch außerhalb der nüchternen Fabrikwelt: IBM Watson oder Alphazero waren genauso „smart“, zu ihrer Zeit. Doch diesmal scheint alles anders. Warum ist das so? Nun, die KI redet jetzt direkt mit uns, jedenfalls tut sie so. Und wir Menschen definieren uns nun einmal eher über Sprache als über die Fähigkeit, mathematische Probleme zu lösen. Daher sind viele von ChatGPT geradezu geschockt. Eine KI, die so sprechen und texten kann, reicht doch sicherlich irgendwie an die Intelligenz von Menschen heran, oder? Man meint, die Sprachroboter bestünden schon das Abitur, und auch das Bestehen des Medizinstudiums scheint nicht mehr weit.

Diese Tatsachen sprechen jedoch eher gegen unsere Art von Prüfungen als für die Intelligenz von Sprachrobotern. Jeder Leser kann die Intelligenz von ChatGPT und ähnlichen Systemen selbst testen. Spiegeln Sie die Antworten der Bots eine Weile ins System zurück. Dann ist nicht mehr viel übrig von der (Sprach-)Intelligenz, jedes sechs Jahre alte Kind würde besser antworten. Infolgedessen stellt sich die sehr wichtige Frage: Wie intelligent ist ChatGPT überhaupt?

Die Intelligenz von Systemen lässt sich beispielsweise in fünf Stufen einteilen: Stufe 1 (Denken und Deduktion), Stufe 2 (Lernen und Induktion), Stufe 3 (Kreativität und Kognition), Stufe 4 (Bewusstsein und Wahrnehmung) und Stufe 5 (Selbstbewusstsein und Selbstwahrnehmung). Viele andere Einteilungen sind möglich, aber an dieser einfachen Einteilung lässt sich gut und unmittelbar erkennen, wo die Sprachsysteme stehen: Sie sind auf Stufe 3 anzusiedeln. Immerhin.

Systeme mit Bewusstsein (Stufe 4) unterscheiden sich prinzipiell von Systemen ohne Bewusstsein. Das merkt man an ihrer Leistungsfähigkeit. Über die Fähigkeit der visuellen Wahrnehmung von Systemen mit Bewusstsein ging es eingangs schon. Aber eine der wichtigsten Eigenschaften bewusster Intelligenzen ist – jedenfalls nach Ansicht des Autors – die der Extrapolation.

Zeigen Sie einem Kind zehn Hunde- und zehn Katzenbilder. Danach wird es Hunde und Katzen für immer unterscheiden können. Zeigen sie der fortschrittlichsten KI zehn Hunde- und zehn Katzenbilder – jeder Fachmann weiß, dass das System hinterher immer noch keine Hunde von Katzen unterscheiden kann. Ein KI benötigt tausend mal mehr Trainingsdaten für die gleiche Aufgabe, natürlich je nach Komplexität. Das muss Gründe haben. Und es hat tiefgreifende Gründe. Ein neuronales Faltungsnetzwerk (CNN), das für solche Aufgaben trainiert wird, hat sehr viele freie Parameter (zum Beispiel repräsentiert durch Gewichtsverbindungen zwischen den einzelnen Neuronen), die man durch den Lernprozess mühsam einstellen muss. Deshalb sind so viele Daten erforderlich, deshalb braucht es „Big Data“. Aber auch ein menschliches Gehirn hat viele freie Parameter, und benötigt dennoch nur extrem wenige Lernbeispiele. Warum? Nun, Systeme mit Bewusstsein nutzen für ihr Agieren in einer komplexen Umgebung die Fähigkeit von Bewusstsein zur erweiterten Generalisierung. Das kann KI (der Stufe 3) so nicht, und sie wird dies auch niemals können.

Nehmen wir ein weiteres Beispiel. Ein Mensch lernt in einem Dorf Autofahren und bekommt nach einer Prüfung und maximal 1000 Kilometern Fahrpraxis seine Lizenz. Am nächsten Tag fährt er in einem anderen Dorf, einer Stadt und sogar in einer Großstadt Auto. Keine KI der Welt kann so etwas leisten, jetzt nicht, und später nicht. KI-Systeme können nur in nahezu ähnlichen Umgebungen fahren, wie die, die vorher eingeübt wurden, das gilt auf den ersten Blick auch für Menschen, aber Menschen können darüber hinaus sehr gut generalisieren und extrapolieren, bis hin zur echten Kreativität. Ja, natürlich können auch KI-Systeme generalisieren, sonst bräuchte man sie nicht – aber sie können es um mehrere Größenordnungen schlechter als wir. Man könnte nun meinen, dass sei eine graduelle Frage, die sich mit mehr Leistungsfähigkeit der Technik weiter verbessern wird. So ist es aber nicht. Menschen sind eben keine (kohlenstoffbasierten) Maschinen, Menschen sind prinzipiell nicht mechanisierbar, also nicht allein mittels physikalischer Gesetze „konstruierbar“ – selbst primitivste Bakterien lassen sich übrigens im Labor nur sehr begrenzt aus komplexen biochemischen Vorbestandteilen synthetisieren.

Stufe-3-Systeme, um im obigen Bild zu bleiben, können Menschen immer (nur) in Spezialfällen überflügeln. Das ist aber nichts Besonderes, sondern zwingend. Ein triviales Beispiel: Ein Bagger kann tiefere Löcher ausheben als ein Mensch, ein Auto fährt schneller als ein Mensch je laufen könnte, dafür ist die Technik schließlich da. Wir bauen Systeme, die uns in jeder ausgewählten Spezialdisziplin überlegen sind – sonst hat Technik überhaupt keinen Sinn. KI-Systeme können besser Schach spielen, besser Go, und nun eben auch gut texten. Das sollte uns alle freuen. Wir Menschen bleiben trotzdem von dieser KI-Stufe 3 der mechanischen Rechner und Texter „unendlich“ weit entfernt, weil Menschen in einer anderen Dimension spielen. Grafiken, die uns auf einer IQ-Skala irgendwo unten platziert zeigen und die KI weit darüber verorten, verleiten die Menschen zu Fehlinterpretationen.

Haben wir uns jemals erschrecken lassen, als wir Grafiken sahen, die zeigten, welche riesigen Kräfte Maschinen im Vergleich zu uns ausüben können. Stufe- 3-Systeme werden uns in mechanisierbarer Intelligenz definitiv übertrumpfen. Man kann sich das so merken: Alles, was irgendwie mathematisierbar ist (so auch der Textfluss von ChatGPT), wird ein KI-System irgendwann einmal besser können als wir. Aber was heißt das schon? Schon ein Taschenrechner kann schneller und besser rechnen als wir selbst dies können.

Heutige KI-Systeme verharren alle auf Stufe 3 – höchstens. Bewusstsein ist mit diesen System nicht erreichbar. Und an die Intelligenzstufe 5, der sich selbst-bewussten Intelligenz, ist für eine heutige KI überhaupt nicht zu denken. Stufe-5-Systeme müssten „sich selbst enthalten“, etwas, was der Mathematiker und Logiker Kurt Gödel bei seinem Unvollständigkeitssatz vor vielen Jahren aufzeigte und nutze. Das menschliche Gehirn und das Selbstbewusstsein sind hochgradig selbstreferenziell, also auf sich selbst beziehend – selbstreferenzielle Hardwaresysteme gibt es gegenwärtig indes nicht. Doch selbst rudimentäres Bewusstsein ist keine Eigenschaft mathematischer Komplexität, und damit von Softwarekomplexität, sondern physikalischer, besser noch chemischer und biologischer Komplexität. Im Rahmen einer sich weiterentwickelnden technischen KI muss also hinreichende physikalische Komplexität gegeben sein. Man muss (für Stufe 4) in sogenannte neuromorphe Systeme investieren, also in Systeme, bei denen die neuronalen Netze physisch existieren, und nicht nur als Software ausgebildet sind. Das ist bei den Sprachrobotern aber nicht der Fall. Alan Turing formulierte im Jahr 1950 einen Test, den berühmten Turing-Test, um die Intelligenz oder die Denkfähigkeit eines Systems mit einem Menschen zu vergleichen, da er die Intelligenz für schwer messbar aber gut vergleichbar hielt. Ein Turing-Test auf Bewusstsein ist daher zwingend und solche Tests lassen sich schon heute skizzieren.

Ein erstes Fazit: Heutige KI-Prozesse sind wie Schattenspiele einer echten (biologischen) Intelligenz. Sie sind prinzipiell nicht in der Lage, in die Intelligenzstufen von bewussten Lebewesen vorzudringen. Wie kommt es, dass manche offenbar glauben, dass ein KI-Gerät der Intelligenzstufe 3 quasi die Weltherrschaft erringen könnte? Wieso warnen selbst sehr renommierte Vertreter der Branche in dieser Weise? Das hat vermutlich viele Gründe und Hintergründe, und einfache Hirngespinste sind es gewiss auch nicht. Die Probleme liegen jedoch nicht in der KI-Technologie, sondern in den Anwendungen, wie beispielsweise potentiellen Eingriffen in Grundrechte durch KI-Richter, Überwachungen auf öffentlichen Plätzen, polizeilichen Verbrechensvorsagen oder militärischen KI-Anwendungen.

Große gesellschaftliche Risiken entstehen, wenn man sich eine voll digitalisierte Welt vorstellt, in der alle Entscheidungen über Menschen immer zuerst im digitalen Universum gefällt werden (müssen): Beförderung, Bestrafung, Gefängnis, Urlaubsbescheid, Nahrungseinkauf, einfach alles. In einer solchen Gesellschaft kann die KI selbstverständlich prinzipiell „die Macht“ übernehmen. Aber letztlich könnte jedes technische Gerät auf diese Weise „Macht ausüben“, wenn man das Gesamtsystem so konstruiert. Das hat nichts mit KI zu tun.

In einer durch und durch digitalisierten Welt mit einer Maschine am Ende der Entscheidungskette haben in Wirklichkeit immer die Eigentümer der Maschinen die Macht. Einfache Bürger würden in einer solchen Gesellschaft keine Chance mehr haben, Entscheidungen einer Maschinen-KI-Bürokratie zu hinterfragen, falls sie das überhaupt noch wollten. Man könnte meinen, so etwas Kafkaeskes würden Menschen niemals bauen. Aber aus dem „Internet der Menschen“ wurde schon das „Internet der Dinge“ – und bald wird es ein „Internet der Körper“ geben, das dürfte jedenfalls nicht mehr aufzuhalten sein.

Von den gegenwärtig verfügbaren KI-Systemen gehen Gefahren aus, die wir auch aus anderen Technologie-Entwicklungen kennen: unbedarfter Einsatz, Gutgläubigkeit der Massen, Machtmissbrauch. Und der Einsatz von Technik lässt seit jeher Fähigkeiten des Menschen verkümmern. Die mechanisierbare Sprach-KI birgt damit das große Risiko, dass nun auch die Sprachfähigkeiten von Menschen massiv reduziert werden, so wie die Rechenfähigkeiten schon enorm reduziert worden sind. Aber ein „Auslöschungsrisiko“ durch eine solche KI anzunehmen, das ist absurd. Weder entsteht eine Superintelligenz, die uns an Intelligenzleistungen übertrifft, noch entsteht eine „Dunkle KI“, welche die Weltherrschaft anstrebt. Wo sollte dieses Streben, der intrinsische Antrieb auch herkommen?

Mit der Beantwortung dieser Frage kommen wir nun zum zweiten und dritten Punkt aus der Einleitung. Nichts spricht dafür, dass physikalische Systeme wie Siliziumkristalle oder Computerchips eine innere Gefühlswelt haben. Stellen wir uns nur einen Moment vor, ein Computerschaltkreis hätte Gefühle. Bei Stromdurchfluss wäre es ihm zu warm, sonst vielleicht zu kalt. Solche Missempfindungen würden dem Computerschaltkreis keinen evolutionären Vorteil bringen, er führt keinen Kampf ums Überleben. Gefühle für unbelebte Systeme sind deshalb nicht nur nicht notwendig, sie sind evolutionär nicht vorgesehen. Nur Systeme, die aufgrund der qualitativen Bewertung einer Wahrnehmung handeln könnten, haben einen Nutzen von Empfindungen. Daran ändern auch Softwarealgorithmen nichts, die auf einem unbelebten Schaltkreis ablaufen.

Da anorganische Chips keine Gefühlswelt besitzen, haben sie auch keine inhärenten Willensprozesse, denn mit dem Willen wird jedes System sich oder seine Umwelt so verändern, dass es zu positiver Wahrnehmung tendiert. Einfach ausgedrückt ist einer anorganischen KI letztlich alles „egal“. Sobald man jedoch biologische KI umsetzt (ein Ziel von Transhumanisten), wird es gefährlich, denn den biologischen Systemen können wir weder inhärente Gefühle noch Willen absprechen. Das wäre daher die Grenzüberschreitung, die wir schon aus anderen Fachgebieten kennen. Sobald der Mensch in Hybris Gott spielt, wird auch dieser Turm zu Babel fallen. Noch ist es aber nicht soweit.

Trotzdem werden bald Vorschläge von den KI-Konzernen präsentiert werden, wie man sich mit ihren Produkten gegen die (angeblich) überbordende Intelligenz der KI wehren kann. Es geht langfristig um nichts Geringeres als die Verschmelzung von Mensch und Maschine. Doch hier sollten wir aufpassen. Lassen wir die Reste unseres Humanismus nicht auch noch transzendieren.

Dr. **Ralf Otte** ist Professor am Institut für Automatisierungssysteme an der Technischen Hochschule Ulm.