Bern

Berner Forscher wollen einen Pakt mit der KI schliessen

Rasante Entwicklung Was, wenn künstliche Intelligenz wie ein Mensch zu denken beginnt? Zwei Wissenschaftler der Uni Bern möchten für diesen Fall einen Vertrag zwischen Menschheit und Maschinen.

Carlo Senn

Federico Benitez (l.) und Walter Senn forschen zum möglichen Bewusstsein von künstlicher Intelligenz. Foto: Franziska Rothenbühler

Der Supercomputer HAL 9000 auf dem Raumschiff hat ein Problem: Er hat einen Fehler gemacht, darum droht ihm die Abschaltung – also sein «Tod». Er wehrt sich mit Gewalt gegen die Besatzung, unterliegt jedoch letztlich einem Astronauten, der ihn abschaltet. Am Ende zeigt der Computer Gefühle: «Ich habe Angst, Dave.»

Eine künstliche Intelligenz, die ein Bewusstsein entwickelt. Das stellte sich Stanley Kubrick bereits in seinem Film «2001: A Space Odyssey» aus dem Jahr 1968 vor.

Ist die künstliche Intelligenz (KI) fast 60 Jahre nach dem Film demnächst an diesem Punkt?

Das glauben zumindest zwei Forscher der Universität Bern: Walter Senn, Professor für Neurowissenschaften und Mathematiker, sowie Federico Benitez, Doktor der Philosophie und Physiker.

Ein Pakt, KI zu kontrollieren

Und die Berner Forscher fordern: Sollte die Menschheit dereinst zur Erkenntnis kommen, dass die von ihnen geschaffenen künstlichen Wesen ein Bewusstsein haben – dann brauche es einen Deal zwischen Mensch und Maschine. Das ungleiche Team hat dazu ein wissenschaftliches Paper verfasst. «Wir möchten verhindern, dass eine Konkurrenz zwischen den Rechten von Menschen und jenen von künstlichen Akteuren entsteht», sagt Benitez. Sie nennen das Abkommen den «Human-AI-Deal».

Wenn die Menschen Vorrang gegenüber den Rechten der künstlichen Agenten haben sollen, obwohl diese ein eigenes Bewusstsein besitzen, müsse den Agenten andere Dinge erlassen werden. Etwa der Schmerz oder zumindest der chronische Schmerz.

Schliesslich seien es ja die Menschen, die die Maschinen erschaffen. Für die Argumentation haben sich Benitez und Senn von Theorien über die Rechte von Tieren inspirieren lassen. Als bewusste Wesen haben auch sie ihre Rechte, wenn auch nicht auf der gleichen Stufe.

Die Idee, den Menschen rechtlich über die Maschine zu stellen, habe insbesondere den Grund, weniger privilegierte Menschen zu schützen. «Unterprivilegierte wären am ehesten von gleichgestellten Maschinen betroffen», sagt Benitez.

Erfahrungsgemäss verstärke zunächst jede neue Technologie die gesellschaftliche Ungleichheit, sagen die Forscher. Deshalb versuchen sie, neben ihrer Forschung über das Hirn und Bewusstsein auch Ideen einzubringen, wie wir als Gesellschaft mit der Möglichkeit von künstlichem Bewusstsein umgehen können.

Baby mit Chip im Hirn

Um das Thema besser zu veranschaulichen, haben die Forschenden ein Gedankenexperiment entwickelt. Sie stellen sich vor, dass einem Säugling mit einer fortschreitenden Hirnkrankheit ein Chip implantiert wird. Dieser ersetzt die geschädigten Hirnareale und kommuniziert mit dem Rest des Hirns und des Körpers. Zusammen wachsen sie, und der Chip übernimmt immer weitere Teile der Hirnfunktion.

Der Chip repräsentiert in diesem Gedankenexperiment eben einen Agenten, der die Hirnfunktionen und wohl auch das Bewusstsein kopiert hat. «Wie könnte man diesem Baby dann absprechen, ein Bewusstsein zu haben?»

Wenn künstliche Agenten dem Menschen ähnlicher werden, hat das Vorteile für beide. Angst zu entwickeln, und später auch Mitgefühl für Menschen, würde für die KI wohl damit beginnen, sich einer existenzbedrohenden Gefahr bewusst zu werden.

«Das Schmerzsystem der Menschen nachzubauen, wäre deshalb wohl sinnvoll», sagt Senn. Gleichzeitig erlaubt die Kontrolle des Schmerzes den erwünschten Spielraum im Umgang mit bewusster künstlicher Intelligenz. Im Handel um rechtliche Privilegien des Menschen könnten den Agenten die eigentlich unnötigen chronischen Schmerzen erspart bleiben.

«Um das friedliche Zusammenleben zu ermöglichen, bräuchte es neben dem Pakt natürlich noch mehr», betont Senn – beispielsweise Regulierungen durch den Staat.

Doch wie weit sind denn eigentlich die aktuellen KI-Systeme?

Eine Studie aus dem Jahr 2023 der Universität Cornell (USA) besagt, dass es zurzeit noch keine KI gibt, die ein Bewusstsein hat – allerdings gebe es keine technischen Barrieren, solche zu entwickeln.

Eine weitere Studie der Universität Bamberg (Deutschland) ist skeptischer: Aus technischen Gründen sei Bewusstsein bei KI-Systemen gegenwärtig nicht möglich.

Bekannt ist, dass Sprachmodelle wie Chat-GPT den Text, der generiert wird, nicht selbst versteht.

Die Beispiele zeigen: Noch sind wir nicht so weit. Dieser Meinung sind auch Senn und Benitez.

Etwas, was einem bewussten Programm aber näher kommen könnte, sind «generative adversariale Netzwerke». Es ist ein KI-Modell, das aus zwei Elementen besteht, die im gegenseitigen Wettbewerb stehen und sich so ständig verbessern.

Die Forscher schlagen vor, dass wir in unserem Hirn ebenfalls solche Gegenspieler haben. Und dass ein weiteres Netzwerk, eben das Bewusstsein, den Wettstreit der Gegenspieler koordiniert und dafür einen Absolutheitsstatus erhalten hat.

Das Dirigenten-Modell

Stellt sich noch die Frage, wann denn eine Maschine bewusst denken würde. Geht es nach den Forschenden, könnte ihr neu entwickeltes Modell darauf Antworten liefern. Denn für Senn als Wissenschaftler sind die Vorgänge im Hirn keine Magie, sondern biochemische Prozesse. Mit dem Modell will Senn das Bewusstsein etwas «entmystifizieren und zugänglicher machen», wie er sagt.

Wenn künstliche Agenten dem Menschen ähnlicher werden, hat das Vorteile für beide.

«Wir denken, dass es im Hirn eine Instanz gibt, die die verschiedenen Informationsströme steuert und abwägt.» Der Dirigent im Hirn. Damit wir in der Welt handeln und bestehen können, muss der Dirigent einen der vielen Informationsströme auswählen, verstärken und umsetzen. Damit wird uns diese Information «bewusst».

Aber wie testet man Bewusstsein? Die Forscher haben sich am Turing-Test orientiert, der durch Fragen und Antworten herausfindet, ob eine Maschine «intelligent» ist. Um den Test auf das Bewusstsein auszuweiten, soll zusätzlich in der «Schaltzentrale» die Funktion eines Dirigenten identifiziert werden.

Der Dirigent, selber nur ein neuronales Netzwerk, steuert gemäss dem Modell im Gehirn drei verschiedene Gruppen von Nervenzellen:

— Neurone, die über Sensoren die Informationen von aussen interpretieren;

— Neurone, die von innen Vorstellungen von Sinneseindrücken generieren; und

— Neurone, die zwischen extern und intern generierten Signalen unterscheiden.

Interessant ist dabei vor allem die dritte Art von Neuronen. Denn deren Aufgabe ist es, zu unterscheiden, was wir als «real» erleben. Der Dirigent kann Neurone verstärken, die uns die Empfindung von «real» vermitteln oder von lediglich «vorgestellt». Mit der dritten Klasse von Neuronen kann er uns aber auch vorgaukeln, dass wir unsere Vorstellung als real empfinden. Genau das geschieht, wenn wir träumen.

Träumen ist entscheidend

«Im Traum sind wir absolut überzeugt, dass wir in der Realität leben.» Neuronen, die uns interne Vorstellungen generieren, sind gleichzeitig aktiv mit den Neuronen, die uns mitteilen, die Signale kämen von aussen. «Nur so können wir mit dem Träumen ein realitätsnahes Verhalten simulieren und erlernen.»

Senn bringt ein Beispiel: Auf einer Safari blicken Sie plötzlich gebannt in die Augen eines Löwen. «Es ist dann günstiger, Sie träumen nachts, wie Sie sich vor dem Löwen knapp ins Auto retten können, als dass Sie aus dem Auto steigen und den Löwen streicheln.»

Senn und Benitez sind der Meinung, dass bei der KI Ähnliches vorgehen könnte. Sie vergleichen das Träumen des Menschen mit dem Trainingsmodus von KI. Das Hirn verarbeitet das Erlebte, lernt Neues, schafft Ordnung.

Anders als Programme mit fixen Abläufen unterscheidet sich die KI unter anderem dadurch, dass sie «lernen» kann. So erkennt eine Bilder-KI dank zahlreicher Trainingsdaten beispielsweise ein Objekt als Tasse.

Es gibt also keinen vorgeschriebenen Ablauf des Programms: Die KI findet durch die neuronalen Netzwerke quasi ihren eigenen Weg und trifft auch eigene Entscheidungen. Auch ein möglicher Dirigent würde lernen, sich anzupassen, tagsüber und im Traum, um die richtigen Informationsströme zu generieren und auszuwählen.

Sollte sich also herausstellen, dass auch die künstlichen Agenten solche Neuronen haben, die eben unterscheiden können zwischen Realität, Vorstellung und Traum, wäre das ein weiteres Indiz, dass diese KI ein Bewusstsein hat. Dafür müsste man in die KI, also in die Software und Hardware hineinschauen können.

Vielleicht werden wir bald von künstlichen Agenten umgeben sein, die von sich behaupten, Bewusstsein zu haben, und die vorgeschlagene Hardware dazu in der Schaltzentrale aufweisen. Dann sei es besser für beide Seiten, jeweils von einem Bewusstsein des Gegenübers auszugehen.

«Wenn wir akzeptieren, dass solche KI-Agenten ein Bewusstsein haben, ist eine friedliche Koexistenz eher möglich», sagt Senn. Damit die Dystopien in den Science-Fiction-Filmen nicht zur Realität werden.

Hinweis: Walter Senn ist ein entfernter Verwandter des Autors.