Unverkäufliche Leseprobe

Stephen Hawking warnt davor, dass die Weiterentwicklung der Künstlichen Intelligenz das Ende der Menschheit bedeute. Andere hingegen feiern das neue Zeitalter der Superintelligenz, da menschliche Kapazitäten mittels intelligenter Maschinen enorm ausgeweitet würden. Sicher ist, dass KI aus dem Bereich der Science-Fiction Einzug in die Realität, ja in unseren Alltag gefunden hat.

Der bekannte Visionär John Brockman hat daher die führenden Wissenschaftler, Philosophen und Künstler unserer Zeit gefragt, was sie von denkenden Maschinen halten.

Der bekannte Visionär *John Brockman*, ehemaliger Aktionskünstler, Herausgeber der Internetzeitschrift »Edge« und Begründer der »Dritten Kultur« (»Third Culture«), leitet eine Literaturagentur in New York und hat bereits zahlreiche Bücher veröffentlicht, u. a. ›Das Wissen von morgen. Was wir für wahr halten, aber nicht beweisen können: Die führenden Wissenschaftler unserer Zeit beschreiben ihre großen Ideen‹, ›Leben, was ist das? Ursprünge, Phänomene und die Zukunft unserer Wirklichkeit‹, ›Welche Idee wird alles verändern? Die führenden Wissenschaftler unserer Zeit über Entdeckungen, die unsere Zukunft bestimmen werden‹, ›Wie funktioniert die Welt? Die führenden Wissenschaftler unserer Zeit stellen die brillantesten Theorien vor‹ und ›Welche wissenschaftliche Idee ist reif für den Ruhestand? Die führenden Köpfe unserer Zeit über die Ideen, die uns am Fortschritt hindern‹.

Weitere Informationen finden Sie auf www.fischerverlage.de

John Brockman

Was sollen wir von Künstlicher Intelligenz halten?

Die führenden Wissenschaftler unserer Zeit über intelligente Maschinen

Aus dem Englischen von Jürgen Schröder

FISCHER Taschenbuch

Erschienen bei FISCHER Taschenbuch
Frankfurt am Main, Juli 2017

Die amerikanische Originalausgabe erschien 2015 unter dem Titel
»What to Think About Machines That Think.
Today's Leading Thinkers on the Age of Machine Intelligence«
im Verlag Harper Collins Publishers, New York

Satz: Dörlemann Satz, Lemförde
Druck und Bindung: CPI books GmbH, Leck
Printed in Germany
ISBN 978-3-596-29705-4

Inhalt

Murray Shannahan

Bewusstsein in der KI auf menschlichem Niveau

Professor für kognitive Robotik am Imperial College, London;
Autor von *Embodiment and the Inner Life*

Nehmen wir einmal an, wir könnten eine Maschine mit einer Intelligenz menschlichen Niveaus ausstatten, das heißt mit der Fähigkeit, einem typischen Menschen in jedem (oder fast jedem) Bereich intellektueller Bemühungen gleichzukommen und vielleicht jeden Menschen in einigen Bereichen zu übertreffen. Hätte eine solche Maschine zwangsläufig Bewusstsein? Das ist eine wichtige Frage, weil eine positive Antwort uns stutzen lassen würde. Wie würden wir mit so etwas umgehen, wenn wir es konstruierten? Wäre es in der Lage, Leid oder Freude zu empfinden? Würde es dieselben Rechte wie ein Mensch verdienen? Sollten wir überhaupt Maschinenbewusstsein erzeugen?

Die Frage, ob eine KI auf Menschenniveau zwangsläufig Bewusstsein hätte, ist außerdem schwierig. Eine Quelle der Schwierigkeit ist die Tatsache, dass mit Bewusstsein bei Menschen und anderen Tieren zahlreiche Attribute verbunden sind. Alle Tiere weisen ein Gefühl für Zwecke auf. Alle (wachen) Tiere sind sich der Welt, in der sie leben und der darin enthaltenden Gegenstände mehr oder weniger bewusst. Alle Tiere zeigen in einem gewissen Grad kognitive Integration, was bedeutet, dass sie alle ihre geistigen Ressourcen – Wahrnehmungen, Erinnerungen und Fertigkeiten – zur Verfolgung ihrer Ziele in der aktuellen Situation in Anschlag bringen können. In dieser Hinsicht weist jedes Tier eine gewisse Art von

Einheit, eine Art von Selbstsein auf. Manche Tiere, einschließlich des Menschen, sind sich auch ihrer selbst bewusst – ihrer Körper und ihres Gedankenflusses. Schließlich sind die meisten, wenn nicht alle Tiere fähig zu leiden, und manche sind in der Lage, Mitgefühl mit dem Leiden anderer zu haben.

Bei (gesunden) Menschen kommen alle diese Attribute als Paket zusammen. Aber in einer KI können sie möglicherweise voneinander getrennt werden. Unsere Frage muss daher verfeinert werden. Welches der Attribute, die wir bei Menschen mit Bewusstsein verbinden, ist, wenn überhaupt, eine notwendige Begleiterscheinung von Intelligenz auf Menschenniveau? Nun, jedes der aufgeführten Attribute (und die Liste ist gewiss nicht erschöpfend) verdient jeweils längere Ausführungen. Ich möchte daher nur zwei herausgreifen – nämlich Bewusstsein der Welt und die Fähigkeit zu leiden. Bewusstsein der Welt, würde ich behaupten, ist tatsächlich eine notwendige Eigenschaft von Intelligenz menschlichen Niveaus.

Sicherlich würde man nicht eine Intelligenz auf menschlichem Niveau zuschreiben, wenn es keine Sprache hätte, und die Hauptverwendung menschlicher Sprache besteht im Sprechen über die Welt. In diesem Sinne ist Intelligenz unzertrennlich mit dem verknüpft, was Philosophen *Intentionalität* nennen. Darüber hinaus ist die Sprache ein soziales Phänomen, und eine grundlegende Verwendung von Sprache innerhalb einer Menschengruppe besteht im Sprechen über Dinge, die alle wahrnehmen können (wie beispielsweise dieses Werkzeug oder jenes Stück Holz), oder wahrgenommen haben (das Stück Holz von gestern) oder wahrnehmen könnten (vielleicht das Stück Holz von morgen). Kurz, die Sprache gründet im Bewusstsein der Welt. Bei einem verkörperten Wesen oder einem Roboter wäre ein solches Bewusstsein aufgrund seiner Interaktionen mit der Umgebung offenbar (Hindernisse vermeiden, Dinge aufnehmen usw.). Aber wir könnten den Begriff auch erweitern, so dass auch eine verteilte, entkörperlichte Künstliche Intelligenz, die mit geeigneten Sensoren ausgestattet ist, darunter fällt.

Um auf überzeugende Weise als eine Facette des Bewusstseins zu gelten, müsste diese Art von Bewusstsein vielleicht Hand in Hand mit einem offenbaren Sinn für Zwecke und einem gewissen Grad kognitiver Integration gehen. Dieses Trio von Attributen wird daher auch bei einer KI als Paket vorhanden sein. Aber lassen wir die Frage einen Augenblick beiseite und kehren wir zur Fähigkeit zurück, Leid und Freude zu empfinden. Im Unterschied zum Bewusstsein der Welt gibt es keinen offensichtlichen Grund für die Annahme, dass Intelligenz menschlichen Niveaus dieses Attribut haben muss, auch wenn es bei Menschen innig mit Bewusstsein verbunden ist. Wir können uns eine Maschine vorstellen, die kalt und gefühllos das ganze Spektrum von Aufgaben ausführt, die beim Menschen Verstand erfordern. Einer solchen Maschine würde dasjenige Attribut des Bewusstseins fehlen, das am meisten zählt, wenn es um das Zugeständnis von Rechten geht. Wie Jeremy Bentham bei der Frage bemerkte, wie man nichtmenschliche Tiere behandeln solle, geht es dabei nicht darum, ob sie räsonieren oder sprechen, sondern ob sie leiden können.

Ich lege hier nicht nahe, dass eine »bloße« Maschine nie fähig sein könnte, Leid oder Freude zu empfinden – dass es in dieser Hinsicht eine biologische Besonderheit gibt. Der Punkt ist vielmehr, dass die Fähigkeit des Empfindens von Leid und Freude von anderen psychologischen Attributen abgekoppelt werden kann, die im menschlichen Bewusstsein zusammengebündelt sind. Aber untersuchen wir diese scheinbare Abkoppelung genauer. Ich stellte bereits die Vorstellung zur Debatte, dass Bewusstsein der Welt mit einem offenkundigen Sinn für Zwecke Hand in Hand gehen könnte. Das Bewusstsein, das ein Tier von der Welt hat, von dem, was ihm die Welt an Gutem oder Schlechtem bietet (in der Ausdrucksweise J. J. Gibsons), dient seinen Zwecken. Ein Tier zeigt das Bewusstsein eines Raubtiers, indem es sich von ihm wegbewegt, und das Bewusstsein einer potentiellen Beute, indem es sich auf sie zubewegt. Vor dem Hintergrund einer Reihe von Zielen und Bedürfnissen ergibt das Verhalten des Tieres einen Sinn. Und

mit Bezug auf einen solchen Hintergrund kann man einem Tier einen Strich durch die Rechnung machen, können seine Ziele unerreicht und seine Bedürfnisse unbefriedigt bleiben. Das ist gewiss die Grundlage für einen Aspekt von Leid.

Wie steht es nun mit Künstlicher Intelligenz menschlichen Niveaus? Würde eine solche KI notwendigerweise eine komplexe Menge von Zielen haben? Könnten ihre Versuche der Zielerreichung nicht vereitelt und an jeder Ecke hintertrieben werden? Wäre es unter diesen scharfen Bedingungen angemessen zu sagen, dass die KI litte, auch wenn ihre Konstitution sie vielleicht immun gegen die Art von Schmerz oder körperlichem Unbehagen machen würde, mit der Menschen vertraut sind?

Hier stößt die Kombination von Einbildungskraft und Intuition an ihre Grenzen. Ich vermute, dass wir nicht herausfinden werden, wie diese Frage zu beantworten sei, bevor wir es nicht mit der echten Sache zu tun haben. Erst wenn eine raffiniertere KI zu einem vertrauten Teil unseres Lebens geworden ist, werden sich unsere Sprachspiele solchen fremdartigen Wesen anpassen. Aber natürlich könnte es bis dorthin zu spät sein, um unsere Meinung darüber zu ändern, ob sie verwirklicht werden sollen. Sie werden wohl oder übel schon da sein.

Steven Pinker

Denken bedeutet nicht Unterjochen

Johnstone-Family-Professor, Abteilung für Psychologie an der Harvard University; Autor von *The Sense of Style: The Thinking Person's Guide to Writing in the Twenty-First Century*

Thomas Hobbes' markige Gleichsetzung von Denken mit »nichts als Rechnen« ist eine der großartigen Ideen der Menschheitsgeschichte. Die Vorstellung, dass Rationalität durch den physischen Prozess des Rechnens erreicht werden kann, wurde im 20. Jahrhundert durch Alan Turings These bestätigt, dass einfache Maschinen jede berechenbare Funktion implementieren können, und durch Modelle von D. O. Hebb, Warren McCulloch und Walter Pitts und ihren wissenschaftlichen Erben, die zeigten, dass Netzwerke aus vereinfachten Neuronen vergleichbare Leistungen erzielen können. Die kognitiven Leistungen des Gehirns lassen sich in physischen Begriffen erklären: Um es grob (und Kritikern ungeachtet) zu formulieren, können wir sagen, dass Überzeugungen eine Art von Informationen sind, das Denken eine Art von Berechnung und Motivation eine Art von Rückkoppelung und Kontrolle.

Aus zwei Gründen ist das eine großartige Idee. Erstens vervollständigt sie ein naturalistisches Verständnis des Universums, indem sie geheimnisvolle Seelen, Geister und Gespenster in der Maschine austreibt. Ebenso wie Darwin es einem nachdenklichen Beobachter der natürlichen Welt ermöglichte, ohne Kreationismus auszukommen, ermöglichten es Turing und andere einem nachdenklichen Beobachter der kognitiven Welt, ohne Spiritualismus auszukommen.

Zweitens öffnet die komputationale Theorie der Vernunft der Künstlichen Intelligenz die Tür – den Denkmaschinen. Ein von Menschen geschaffener Informationsprozessor könnte die Kräfte des menschlichen Geistes im Prinzip kopieren und übertreffen. Nicht, dass das wahrscheinlich auch in der Praxis geschehen wird, da wir wahrscheinlich nie die anhaltende technische und wirtschaftliche Motivation erleben werden, die dafür notwendig ist. Ebenso wie die Erfindung des Autos keine Verdopplung des Pferdes erforderte, würde die Entwicklung eines KI-Systems, das sich selbst unterhalten könnte, keine Verdopplung eines Exemplars von *Homo sapiens* erfordern. Ein Gerät, das dazu entworfen wurde, Auto zu fahren oder eine Epidemie vorherzusagen, muss nicht dazu entworfen sein, einen Paarungspartner anzuziehen oder verfaultes Aas zu meiden.

Dennoch haben vor kurzem erzielte Minischritte in Richtung auf intelligentere Maschinen zu einer Wiederbelebung der wiederkehrenden Angst geführt, dass unsere Erkenntnis für unseren Untergang verantwortlich sein wird. Meine eigene Ansicht ist, dass gegenwärtige Ängste mit Bezug auf amoklaufende Computer eine Verschwendung emotionaler Energie sind – dass das Szenario mehr Ähnlichkeit mit dem Jahr-2000-Problem als mit dem Manhattan-Projekt hat.

Denn zum einen steht uns eine lange Zeit zur Verfügung, um uns darauf vorzubereiten. KI auf menschlichem Niveau ist immer noch der fünfzehn bis zwanzig Jahre entfernte Standard, so wie sie es immer gewesen ist, und viele ihrer in jüngster Zeit angepriesenen Fortschritte haben flachgründige Wurzeln. Es stimmt zwar, dass »Experten« in der Vergangenheit auf groteske Weise die Möglichkeit technischer Fortschritte, die sich schnell vollzogen, von der Hand gewiesen haben. Aber das ist zweischneidig: »Experten« haben auch unmittelbar bevorstehende Fortschritte angekündigt (oder sind darüber in Panik geraten), die nie eintraten, wie zum Beispiel mit Kernenergie betriebene Autos, Unterwasserstädte, Kolonien auf dem Mars, Designerbabys und Lagerhäuser mit Zombies, die

am Leben erhalten werden, um Menschen mit Ersatzorganen zu versorgen.

Außerdem ist es absonderlich zu meinen, dass die Robotik-Ingenieure bei ihrer Arbeit keine Schutzmaßnahmen gegen Unheil einbauen würden. Dafür würden sie keine schwergewichtigen »Regeln der Robotik« oder irgendeine neumodische Moralphilosophie brauchen, sondern nur denselben gesunden Menschenverstand, der in den Entwurf von Nahrungsprozessoren, Tischsägen, Heizgeräten und Autos einfloss. Die Sorge, dass ein KI-System so klug mit Bezug auf das Erreichen eines seiner programmierten Ziele (wie die Beschaffung von Energie) werden würde, dass es die anderen (wie menschliche Sicherheit) rücksichtslos überginge, nimmt an, dass die KI schneller über uns hereinbrechen wird, als wir ausfallsichere Vorsichtsmaßnahmen entwerfen können. Die Wirklichkeit ist, dass der Fortschritt in der KI dem Medienrummel zum Trotz langsam ist, und es wird viel Zeit für Feedback von schrittweisen Implementationen geben, wobei die Menschen in jedem Stadium den Schraubenzieher in der Hand halten.

Würde ein künstlich intelligentes System diese Schutzmaßnahmen *vorsätzlich* deaktivieren? Warum würde es das wollen? Die Dystopien der KI projizieren eine engstirnige Alpha-Männchen-Psychologie auf den Begriff der Intelligenz. Sie nehmen an, dass übermenschlich intelligente Roboter Ziele entwickeln würden wie die Absetzung ihrer Herren oder die Übernahme der Weltmacht. Aber Intelligenz ist die Fähigkeit, neue Mittel zu ersinnen, um ein Ziel zu erreichen; die Ziele sind der Intelligenz selbst äußerlich. Klug zu sein ist nicht dasselbe wie etwas zu wollen. Die Geschichte wartet mit dem gelegentlichen größenwahnsinnigen Despoten oder dem psychopathischen Serienmörder auf, aber diese sind die Erzeugnisse einer Geschichte natürlicher Selektion, die testosteronsensitive Schaltkreise bei einer bestimmten Primatenspezies formt, und kein unvermeidliches Merkmal intelligenter Systeme. Es ist bezeichnend, dass viele unserer Technopropheten gar nicht die Möglichkeit erwägen, dass die KI sich entlang

weiblicher Vorgaben entwickeln wird – indem sie völlig in der Lage ist, Probleme zu lösen, aber ohne den Wunsch, Unschuldige zu vernichten oder die Kultur zu beherrschen.

Wir können uns einen maliziösen *Menschen* vorstellen, der ein Bataillon von Robotern entwirft und freisetzt, um Massenvernichtung zu säen. Aber Katastrophenszenarios lassen sich in der Vorstellung billig durchspielen, und wir sollten die Kette von Wahrscheinlichkeiten im Auge behalten, die sich entfalten müsste, bevor dieses Szenario Wirklichkeit würde. Es müsste ein böser Geist aufsteigen, der sowohl einen Hunger nach sinnlosem Massenmord als auch Brillanz in technischen Erfindungen besitzt. Er müsste ein Team von Mitverschwörern rekrutieren und verwalten, die vollkommene Geheimhaltung, Loyalität und Kompetenz ausübten. Und die Operation müsste die Gefahren des Entdecktwerdens, des Verrats, von Sting-Operationen, grober Fehler und Missgeschicke überleben. Theoretisch könnte das passieren, aber es gibt dringlichere Dinge, über die wir uns Sorgen machen sollten.

Wenn wir die Science-Fiction-Katastrophenszenarien beiseitelegen, ist die Möglichkeit einer hochentwickelten Künstlichen Intelligenz erhebend – nicht nur im Hinblick auf die praktischen Vorteile, wie beispielsweise die phantastischen Gewinne an Sicherheit, Freizeit und Umweltfreundlichkeit selbstgesteuerter Autos, sondern auch im Hinblick auf die philosophischen Möglichkeiten. Die komputationale Theorie des Geistes hat niemals die Existenz von Bewusstsein im Sinne der Subjektivität in der ersten Person erklärt (obwohl sie vollkommen in der Lage ist, die Existenz von Bewusstsein im Sinne von zugänglicher und berichtbarer Information zu erklären). Ein Vorschlag lautet, dass Subjektivität von Natur aus jedem hinreichend komplizierten kybernetischen System zukommt. Ich pflegte zu meinen, dass diese Hypothese (ebenso wie ihre Alternativen) sich beständig nicht überprüfen ließe. Aber stellen wir uns einen intelligenten Roboter vor, der darauf programmiert ist, seine eigenen Systeme zu überwa-

chen und wissenschaftliche Fragen zu stellen. Wenn er, ohne dass er dazu aufgefordert wird, die Frage stellen würde, warum er selbst subjektive Erlebnisse hat, würde ich den Gedanken ernst nehmen.