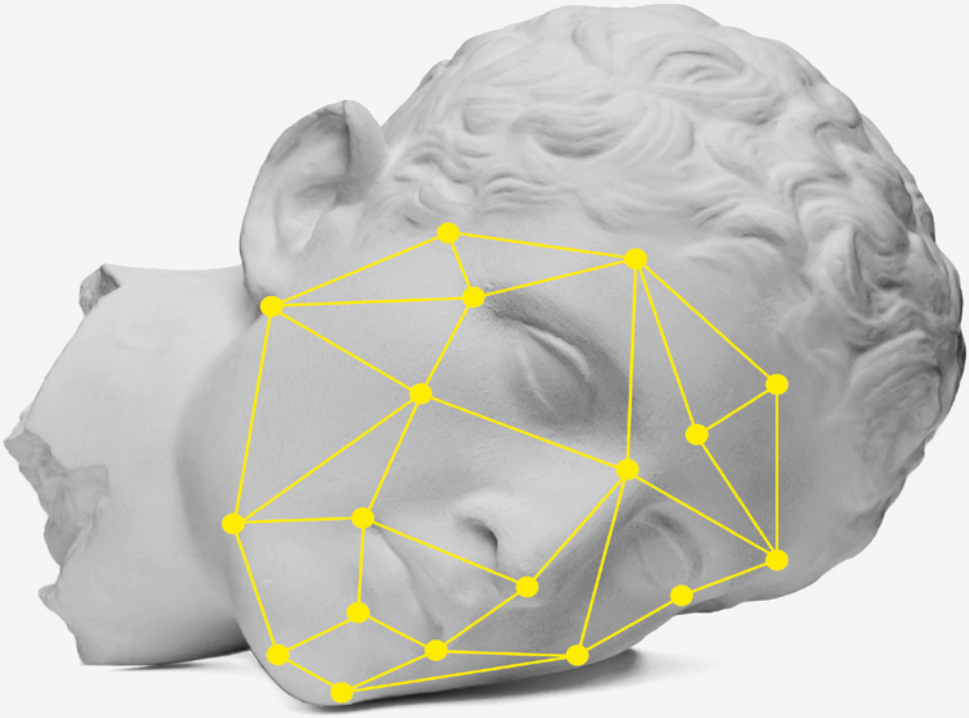
STUART RUSSELL

HUMAN COMPATIBLE



Künstliche Intelligenz

und wie der Mensch die Kontrolle über superintelligente Maschinen behält

mitp

INHALT

VORWORT

Warum Sie dieses Buch genau jetzt lesen sollten

In diesem Buch geht es um die Vergangenheit, die Gegenwart und die Zukunft unserer Bestrebungen, Intelligenz zu verstehen und zu erschaffen. Das ist ein wichtiges Thema, denn die KI (*Künstliche Intelligenz*, auch AI für *Artificial Intelligence*) ist unserer Tage nicht nur bereits allgegenwärtig, sondern wird auch und vor allem die wohl wichtigste Technologie der Zukunft sein. Die Weltmächte werden sich dessen langsam bewusst – und die größten Unternehmen wissen es schon seit geraumer Zeit. Wir können nicht genau sagen, wie und mit welchem Tempo sich diese Technologie weiterentwickeln wird. Aber eines ist klar: Wir müssen uns darauf vorbereiten, dass Maschinen die menschliche Fähigkeit zur Entscheidungsfindung eines Tages vielleicht übertreffen werden. Welche Folgen hätte das?

Alles, was unsere Zivilisation hervorgebracht hat, ist das Ergebnis unserer Intelligenz. Wenn wir also Zugang zu einer deutlich höheren Intelligenz hätten, wäre dies zweifellos das größte Ereignis der Menschheitsgeschichte. In diesem Buch wird erklärt, warum es sich dabei auch um das letzte Ereignis handeln könnte – und wie wir das verhindern können.

Was Sie erwartet

Dieses Buch ist in drei große Themenbereiche unterteilt: Der erste Teil in den Kapiteln 1 bis 3 beschäftigt sich mit menschlicher und maschineller Intelligenz an sich. Dabei wird keine technische Vorbildung vorausgesetzt. Wenn Sie tiefer in die Materie einsteigen wollen, finden Sie eine Erläuterung der Grundkonzepte moderner KI-Systeme in den Anhängen A bis D. Der zweite Teil, Kapitel 4 bis 6, behandelt einige der Probleme, die aus intelligenten Maschinen erwachsen. Das größte davon ist, wie wir es schaffen können, jederzeit die Kontrolle über Maschinen zu behalten, die mächtiger sind als wir. Im dritten Teil in den Kapiteln 7 bis 10 wird ein neuer Ansatz der KI-Forschung vorgestellt, der garantieren soll, dass die Maschinen dem Menschen

auf alle Zeit nützlich und dienlich sind. Das Buch richtet sich an alle interessierten Leserinnen und Leser. Ich hoffe allerdings, dass es auch von vielen Spezialisten der künstlichen Intelligenz gelesen wird und sie dazu anregt, ihre grundlegenden Anschauungen zu überdenken.

DANKSAGUNG

An der Entstehung dieses Buchs waren viele Menschen beteiligt. Dazu gehören Paul Slovak, Lektor bei Viking, sowie Laura Stickney, Lektorin bei Penguin, mein Literaturagent John Brockman, der mich zum Schreiben ermutigt hat, Jill Leovy und Rob Reid, die ein Füllhorn an nützlichem Feedback für mich hatten, und weitere Leser der ersten Entwürfe, insbesondere Ziyad Marar, Nick Hay, Toby Ord, David Duvenaud, Max Tegmark und Grace Cassy. Caroline Jeanmaire war eine unschätzbare Hilfe beim Sammeln und Zusammenstellen der unzähligen Verbesserungsvorschläge der Testleser und Martin Fukui hat sich für mich um die Bildrechte gekümmert.

Die wesentlichen technischen Konzepte in diesem Buch habe ich gemeinsam mit Mitgliedern des *Center for Human-Compatible AI* in Berkeley entwickelt. Besonders erwähnen möchte ich Tom Griffiths, Anca Dragan, Andrew Critch, Dylan Hadfield-Menell, Rohin Shah und Smitha Milli. Das Center wird auf bewundernswerte Weise von Mark Nitzberg und Rosie Campbell geleitet und von der Open Philanthropy Foundation großzügig mit finanziellen Mitteln ausgestattet.

Ramona Alvarez und Carine Verdeau haben dafür gesorgt, dass während der gesamten Entstehungsgeschichte des Buchs alles glatt vonstattenging. Meine unglaubliche Ehefrau Loy und unsere gemeinsamen Kinder Gordon, Lucy, George und Isaac haben mit ihrer unermesslichen Liebe, Nachsicht und Unterstützung zum Gelingen und Abschluss dieses Projekts beigetragen.

1

WENN WIR ERFOLG HABEN

Vor längerer Zeit lebten meine Eltern im englischen Birmingham in einem Haus in der Nähe der Universität. Sie entschieden sich, von der Stadt aufs Land zu ziehen, und verkauften das Haus an David Lodge, einen Professor für englische Literatur. Lodge war zu jener Zeit ein bekannter Romanautor. Ich habe ihn zwar nie persönlich kennengelernt, entschloss mich aber, einige seiner Werke zu lesen: *Ortswechsel* (Originaltitel: *Changing Places*) und *Schnitzeljagd* (Originaltitel: *Small World*). Zu den Hauptakteuren gehören fiktionale Akademiker, die aus einem fiktionalen Birmingham in ein fiktionales Berkeley in Kalifornien ziehen. Da ich ein echter Akademiker aus dem echten Birmingham bin, der gerade ins echte Berkeley gezogen war, rief mich das Ministerium für Zufälle offenbar dazu auf, genauer hinzusehen.

Eine Szene aus *Schnitzeljagd* hat es mir besonders angetan: Der Protagonist, ein aufstrebender Literaturtheoretiker, besucht eine internationale Konferenz und fragt ein hochkarätig besetztes Podium: »Was geschieht, wenn alle Ihnen zustimmen?« Die Frage sorgt für Verblüffung, denn es ging den Diskutanten mehr um den geistigen Schlagabtausch und weniger um Wahrheitsfindung oder Erkenntnisgewinn. In dem Moment wurde mir klar, dass man den führenden Persönlichkeiten in der KI eine ganz ähnliche Frage stellen könnte: »Was, wenn Sie Erfolg haben?« Diese Disziplin hat sich stets das Ziel gesetzt, eine KI mit menschlichen oder gar übermenschlichen Fähig-

keiten zu erschaffen. Was für Folgen das haben würde, hat man sich allerdings nur selten oder gar nicht gefragt.

Ein paar Jahre später begannen Peter Norvig und ich mit der Arbeit an einem neuen KI-Lehrbuch, das erstmals 1995 veröffentlicht wurde.[1] Der letzte Abschnitt mit dem Titel »What If We Do Succeed?« (Was, wenn wir Erfolg haben?) weist auf mögliche positive und negative Folgen hin, ohne ein eindeutiges Fazit zu ziehen. Als 2010 die dritte Auflage erschien, waren viele Menschen zu dem Schluss gekommen, dass eine übermenschliche KI vielleicht gar keine so gute Sache wäre. Allerdings waren die meisten von ihnen eher fachfremd und gehörten nicht zum Mainstream der KI-Forscher. 2013 war ich inzwischen davon überzeugt, dass dieses Thema nicht nur mitten in die KI-Forschung gehört, sondern möglicherweise die wichtigste Frage für die gesamte Menschheit darstellt.

Im November 2013 hielt ich einen Vortrag in der Dulwich Picture Gallery, einem ehrwürdigen Kunstmuseum im Süden Londons. Das Publikum bestand größtenteils aus interessierten Seniorinnen und Senioren ohne wissenschaftliche Vorbildung. Ich musste bei meinem Vortrag also komplett auf technische Fachbegriffe verzichten. Das war eine hervorragende Gelegenheit, meine Überlegungen erstmals in der Öffentlichkeit zu präsentieren. Nachdem ich erklärt hatte, worum es bei KI geht, nominierte ich fünf Kandidaten für das »größte Ereignis in der Zukunft der Menschheit«:

1. Wir sterben aus (Asteroidenaufprall, Klimakatastrophe, Pandemie usw.).
2. Wir leben ewig (medizinische Lösung gegen das Altern).
3. Wir finden eine Möglichkeit, schneller als das Licht zu reisen, und erobern das Universum.
4. Wir erhalten Besuch von einer uns überlegenen außerirdischen Zivilisation.
5. Wir erfinden eine superintelligente KI.

Ich votierte für die fünfte Option, die superintelligente KI, denn diese würde uns helfen, (Natur-)Katastrophen zu vermeiden, den Schlüssel zum ewigen Leben zu finden und schneller als das Licht zu reisen – sofern das überhaupt möglich ist. Das wäre wirklich ein gewaltiger Sprung für unsere Zivilisation. Danach wäre nichts mehr wie zuvor. Eine superintelligente KI würde in vielen Punkten der Ankunft

einer überlegenen Alien-Rasse ähneln, ist aber sehr viel wahrscheinlicher. Und was noch wichtiger ist: Bei der KI haben wir ein Wörtchen mitzureden, bei den Außerirdischen wohl eher nicht …

Dann fragte ich das Publikum, was wohl geschehen würde, wenn wir heute eine Botschaft von Außerirdischen erhielten, die deren Ankunft auf der Erde in 30 Jahren ankündigt. Das Wort »Hölle« dürfte für das zu erwartende Chaos kaum ausreichen. Unsere Reaktion auf die zu erwartende Ankunft von superintelligenter KI fällt dagegen mehr als verhalten aus. (In einem späteren Vortrag habe ich das durch den in Abbildung 1.1 dargestellten E-Mail-Austausch verdeutlicht.) Abschließend habe ich die Bedeutung einer superintelligenten KI wie folgt erklärt: »Ein Erfolg wäre das größte Ereignis in der Menschheitsgeschichte … aber vielleicht auch das letzte.«

Von: Überlegene außerirdische Zivilisation <sac12@sirius.canismajor.u>
An: Menschheit@UN.org
Betreff: Kontakt
Achtung: Wir kommen in 30 bis 50 Jahren an

Von: Menschheit@UN.org
An: Überlegene außerirdische Zivilisation <sac12@sirius.canismajor.u>
Betreff: Abwesenheitsbenachrichtigung: Re: Kontakt
Die Menschheit ist gerade nicht erreichbar. Wir melden uns, wenn wir wieder im Büro sind. ☺

Abb. 1.1: Vermutlich nicht die Antwort, die wir einer überlegenen Zivilisation von Außerirdischen nach deren Kontaktaufnahme senden würden

Ein paar Monate später, im April 2014, besuchte ich eine Konferenz in Island. National Public Radio rief mich an und bat um ein Interview zum gerade in den Vereinigten Staaten gestarteten Kinofilm *Transcendence*. Ich hatte zwar Zusammenfassungen der Handlung und auch einige Kritiken gelesen, den Film aber nicht gesehen, denn in meiner damaligen Wahlheimat Paris sollte er erst im Juni in die Kinos kommen. Allerdings flog ich auf dem Rückweg erst nach Boston, da ich dort an einer Besprechung des Verteidigungsministeriums teilnehmen wollte. Nach meiner Ankunft am Bostoner Logan Airport nahm ich ein Taxi zum nächsten Kino, in dem der Film lief. In der zweiten Reihe sitzend, verfolgte ich, wie ein von Johnny Depp gespielter KI-Professor der

Uni Berkeley von fanatischen KI-Skeptikern niedergeschossen wird, die sich vor einer superintelligenten KI fürchten. Unwillkürlich sank ich in meinem Sitz zusammen. (Hatte sich hier wieder einmal das Ministerium für Zufälle bei mir gemeldet?) Bevor der von Johnny Depp gespielte Wissenschaftler stirbt, wird sein Geist in einen Quanten-Supercomputer übertragen, übertrifft bald die menschlichen Fähigkeiten und droht, die Weltherrschaft an sich zu reißen.

Am 19. April 2014 erschien in der *Huffington Post* eine von den Physikern Max Tegmark, Frank Wilczek und Stephen Hawking gemeinsam verfasste Besprechung zu *Transcendence*. Sie enthielt den Satz über das größte Ereignis in der Menschheitsgeschichte, den ich in meiner Rede in Dulwich verwendet hatte. Seither gelte ich in der Öffentlichkeit als der Forscher, der sein eigenes Forschungsgebiet für eine potenzielle Gefahr für uns als Spezies hält.

Was bisher geschah ...

Die Anfänge der KI reichen zurück bis in die Antike, aber der »offizielle« Startschuss fiel 1956. Die beiden jungen Mathematiker John McCarthy und Marvin Minsky hatten den bereits als Erfinder der Informationstheorie berühmt gewordenen Claude Shannon sowie Nathaniel Rochester, den Entwickler des ersten kommerziellen Computers von IBM, dazu überredet, mit ihnen gemeinsam eine Sommerschule am Dartmouth College zu organisieren. Das Ziel wurde wie folgt beschrieben:

> »Die Studie soll von der Annahme ausgehen, dass grundsätzlich alle Aspekte des Lernens und anderer Merkmale der Intelligenz so genau beschrieben werden können, dass eine Maschine dazu gebracht werden kann, sie zu simulieren. Es soll versucht werden, herauszufinden, wie Maschinen dazu gebracht werden können, Sprache zu benutzen, Abstraktionen und Begriffe zu bilden, Probleme zu lösen, die zu lösen bislang dem Menschen vorbehalten sind, und sich selbst zu verbessern. Wir denken, dass ein bedeutender Fortschritt auf einem oder mehreren dieser Gebiete erzielt werden kann, wenn eine sorgfältig ausgewählte Gruppe von Wissenschaftlern einen Sommer lang zusammen daran arbeitet.«

Wir wissen heute, dass der Sommer nicht ausgereicht hat. Tatsächlich arbeiten wir noch heute an all diesen Problemen.

Im ersten Jahrzehnt nach der Dartmouth-Konferenz konnte die KI mehrere große Erfolge verbuchen, darunter den Algorithmus von Alan Robinson für allgemeine logische Schlussfolgerungen[2] und das Dameprogramm von Arthur Samuel, das sich selbst beibrachte, seinen Schöpfer zu schlagen.[3] Die erste KI-Blase platzte in den späten 1960er-Jahren, da frühe Bemühungen in den Bereichen Machine Learning und maschinelle Übersetzung die in sie gesetzten Erwartungen nicht erfüllten. Ein von der britischen Regierung 1973 in Auftrag gegebener Bericht kam zu folgendem Schluss: »In keinem Bereich des Forschungsgebiets haben die bisherigen Entdeckungen zu den versprochenen gewaltigen Auswirkungen geführt.«[4] Anders ausgedrückt: Die Maschinen waren einfach nicht klug genug.

Mit meinen elf Jahren kannte ich diesen Bericht zum Glück nicht. Als ich zwei Jahre später den programmierbaren Taschenrechner Sinclair Cambridge bekam, wollte ich ihn intelligent machen. Leider konnten Programme für das Gerät maximal 36 Zeichen umfassen, was für KI auf menschlichem Level nicht ausreicht. Das konnte mich nicht aufhalten. Nachdem ich Zugang zu dem riesigen Supercomputer CDC 6600[5] am Imperial College London bekommen hatte, schrieb ich ein Schachprogramm: einen 60 cm hohen Lochkartenstapel. Das Programm selbst war nicht besonders gut, aber das war nicht wichtig. Ich kannte jetzt meine Bestimmung.

Mitte der 1980er war ich Professor in Berkeley und die KI war wieder im Kommen – dank des kommerziellen Potenzials der sogenannten Expertensysteme. Die zweite KI-Blase platzte, als sich diese Systeme für viele der Aufgaben, die sie übernehmen sollten, als unzureichend erwiesen. Auch hier waren die Maschinen nicht klug genug. Ein KI-Winter brach an. Mein eigener KI-Kurs in Berkeley, der heute mit über 900 Studierenden aus allen Nähten platzt, wurde 1990 von lediglich 25 Personen besucht.

Die KI-Community hatte ihre Lektion gelernt: Klüger war in jedem Fall besser. Aber wie ließ sich das erreichen? Die Mathematik sollte es richten. Wir knüpften an bewährte Disziplinen an: Wahrscheinlichkeitsrechnung, Statistik und Kontrolltheorie. Die Saat für die heutigen Fortschritte wurde in jenem KI-Winter gelegt. Dazu gehörten frühe Arbeiten an großen probabilistischen Schlussfolgerungssystemen und

einem Gebiet, das später unter der Bezeichnung *Deep Learning* bekannt wurde.

Ab etwa 2011 ermöglichten Deep-Learning-Techniken dramatische Durchbrüche in der Spracherkennung, der visuellen Objekterkennung und der maschinellen Übersetzung: drei der wichtigsten bisher ungelösten Probleme der KI. Bei einigen Tests bewältigen Maschinen diese Aufgaben heute ebenso gut oder besser als der Mensch. In den Jahren 2016 und 2017 besiegte AlphaGo von DeepMind den ehemaligen Go-Weltmeister Lee Sedol und den amtierenden Meister Ke Jie. Das hatten Fachleute frühestens für das Jahr 2097 erwartet, wenn überhaupt jemals.[6]

Heute lesen wir in den Medien fast täglich von den Fortschritten der KI. Tausende von Start-ups wurden gegründet, von reichlich Risikokapital finanziert. Millionen Studierende absolvieren Onlinekurse zu KI und Machine Learning. Kein Wunder, locken doch Spitzengehälter von mehreren Millionen Dollar. Die Investitionen von Venture-Fonds, Regierungen und Großunternehmen in diesem Bereich betragen zig Milliarden Dollar pro Jahr. Allein in den letzten fünf Jahren wurde das Feld besser mit Finanzmitteln ausgestattet als in seiner gesamten bisherigen Geschichte. Die Projekte, an denen gerade gearbeitet wird, werden im Laufe des nächsten Jahrzehnts höchstwahrscheinlich große Auswirkungen auf unser Leben haben: selbstfahrende Autos zum Beispiel oder intelligente persönliche Assistenten. Die möglichen wirtschaftlichen und sozialen Vorteile der KI sind gewaltig. Das sorgt natürlich für enormen Schwung in der KI-Forschung.

... und was uns noch erwartet

Bedeutet dieser rasante Fortschritt, dass die Maschinen dabei sind, uns zu überholen? Keineswegs. Bevor es superintelligente Maschinen geben wird, sind noch einige Hürden zu überwinden.

Wissenschaftliche Durchbrüche sind generell kaum vorhersagbar. Wie komplex das Thema ist, zeigt die Geschichte eines anderen Forschungsgebiets, das ebenfalls das Potenzial hat, der Zivilisation ein Ende zu bereiten: die Kernphysik.

Anfang des 20. Jahrhunderts gab es wohl keinen bekannteren Kernphysiker als Ernest Rutherford, den Entdecker des Protons, den Mann, der das erste Atom spaltete (siehe Abbildung 1.2 [a]). Wie seine

Kollegen war sich Rutherford bereits seit langer Zeit darüber im Klaren, dass Atomkerne gewaltige Mengen an Energie speichern. Allerdings war die vorherrschende Meinung, dass man diese Energiequelle nicht anzapfen könne.

Am 11. September 1933 hielt die britische Wissenschaftliche Gesellschaft (*British Association for the Advancement of Science*) in Leicester ihre Jahrestagung ab. Lord Rutherford war Sprecher der abendlichen Sitzung. Wie schon zu früheren Gelegenheiten machte er den Anwesenden die Kernenergie betreffend keine großen Hoffnungen: »Jeder, der in der Umwandlung dieser Atome eine Energiequelle sieht, redet Unsinn.«[7] Am nächsten Morgen war Rutherfords Rede in der Londoner *Times* abgedruckt (siehe Abbildung 1.2 [b]).

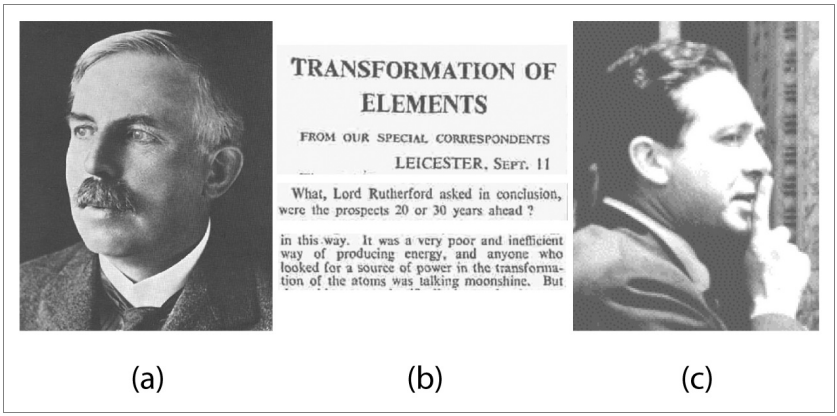


Abb. 1.2: (a) Lord Rutherford, Kernphysiker, (b) Auszug aus einem Bericht der Times vom 12. September 1933 über eine Rede, die Rutherford am Vorabend gehalten hatte, (c) Leó Szilárd, Kernphysiker

Leó Szilárd (siehe Abbildung 1.2 [c]), ein ungarischer Physiker, der kurz zuvor vor dem Nationalsozialismus aus Deutschland geflohen war, wohnte zu der Zeit im Imperial Hotel am Russell Square in London. Beim Frühstück las er den Artikel in der *Times*. Über das Gelesene nachsinnend, unternahm er einen Spaziergang, bei dem ihm die Idee zu der durch Neutronen hervorgerufenen nuklearen Kettenreaktion kam.[8] So wurde das »unmöglich lösbare« Problem der Nutzung der Kernenergie praktisch in weniger als 24 Stunden gelöst. Im folgenden Jahr reichte Szilárd ein geheimes Patent für einen Kernreaktor ein. Das erste Patent für eine Atomwaffe wurde 1939 in Frankreich erteilt.

Und die Moral der Geschichte? Eine Wette gegen den menschlichen Einfallsreichtum ist leichtsinnig, vor allem wenn es um unser aller Zukunft geht. In der KI-Community macht sich eine Art Verleugnungshaltung breit: Manchmal wird bereits die Möglichkeit eines Erfolgs der langfristigen KI-Ziele bestritten. Das erinnert an einen Busfahrer, der die gesamte Menschheit herumkutschiert. In einem Affenzahn hält er auf eine Klippe zu und behauptet gleichzeitig: »Ja klar stürzen wir bei dem Tempo über die Klippe. Aber keine Angst, uns geht vorher noch das Benzin aus!«

Ich sage keinesfalls, dass der Erfolg in der KI *unabdingbar* ist. Und ich bin mir ziemlich sicher, dass es in den nächsten Jahren nicht dazu kommen wird. Dennoch ist es sicher klug, sich auf die Möglichkeit vorzubereiten. Wenn alles gut geht, brechen goldene Zeiten für die Menschheit an, aber wir müssen uns vor Augen halten, dass wir an Dingen arbeiten, die sehr viel mächtiger sind als der Mensch. Wie können wir denn sicherstellen, dass diese Dinge niemals in irgendeiner Weise Macht über uns haben werden?

Damit Sie besser verstehen, wie gefährlich das Ganze ist, hilft ein Blick auf die Auswahl- und Empfehlungsalgorithmen in den sozialen Medien. Sie sind nicht besonders intelligent, und doch beeinflussen sie Milliarden von Menschen und damit indirekt die ganze Welt. Üblicherweise zielen solche Algorithmen darauf ab, die *Klickrate* zu maximieren. Die Klickrate bezeichnet die Wahrscheinlichkeit, mit der ein Nutzer auf die präsentierten Elemente klickt. Wenn wir also einen maximalen Wert erzielen möchten, müssen wir nur solche Elemente anzeigen, auf die Nutzer gern klicken, richtig? Falsch. Die Lösung besteht darin, die Vorlieben der Nutzer so zu verändern, dass sie transparenter werden. Einem besser durchschaubaren Nutzer können Elemente präsentiert werden, auf die er wahrscheinlich klickt und somit mehr Einnahmen generiert. Menschen mit sehr extremen politischen Ansichten sind in dieser Hinsicht meist besser durchschaubar. (Möglicherweise gibt es auch Artikel, auf die extrem gemäßigte Politikinteressierte mit hoher Wahrscheinlichkeit klicken, aber das Thema dieser Artikel ist schwer vorstellbar.) Wie ein rationales Wesen lernt der Algorithmus, wie er den Zustand seiner Umgebung verändern kann, um die eigene Belohnung zu maximieren.[9] In diesem Fall ist die Umgebung die Meinung des Nutzers. Die Folgen sind ein Wiederaufleben des Faschismus, die Kündigung des Gesellschaftsvertrags, auf dem

Demokratien weltweit gründen, und möglicherweise das Ende der Europäischen Union und der NATO. Erstaunlich, was ein paar Zeilen Programmcode anrichten können. Wenn schon ein relativ einfacher Algorithmus dazu imstande ist, wozu ist dann erst ein *wirklich* intelligenter Algorithmus in der Lage?

Was lief schief?

Die Geschichte der KI wird von einem zentralen Mantra bestimmt: »Je intelligenter, desto besser.« Ich bin überzeugt, dass das ein Fehler ist, und zwar nicht, weil ich irgendwie befürchte, ersetzt zu werden, sondern aufgrund unseres Verständnisses von Intelligenz an sich.

Unsere Vorstellung von Intelligenz ist zentral für unser Selbstverständnis. Nicht umsonst bezeichnen wir uns als *Homo sapiens*, als »weiser Mensch«. Nach über 2.000 Jahren der Selbstreflexion lässt sich unsere Idee von Intelligenz wie folgt zusammenfassen:

> Menschen sind insoweit intelligent, als unsere Handlungen darauf ausgerichtet sind, unsere Ziele zu erreichen.

Alle anderen Merkmale für Intelligenz – wahrnehmend, denkend, lernend, erfindend und so weiter – lassen sich als Bestandteile unseres erfolgreichen Handlungsvermögens betrachten. Seit Anbeginn der KI-Forschung wurde Intelligenz in Maschinen ebenso definiert:

> Maschinen sind insoweit intelligent, als ihre Handlungen darauf ausgerichtet sind, ihre Ziele zu erreichen.

Doch weil Maschinen, anders als wir Menschen, keine eigenen Ziele haben, geben wir ihnen diese Ziele vor. Mit anderen Worten: Wir entwickeln Maschinen, die etwas optimieren sollen, geben ihnen Ziele vor und drücken den Startknopf.

Diesen Ansatz finden wir nicht nur auf dem Gebiet der KI. Er zieht sich wie ein roter Faden durch die technologischen und mathematischen Grundsteine unserer Gesellschaft. In der Kontrolltheorie werden Steuerungen für alles Mögliche entwickelt – vom Jumbojet bis zur Insulinpumpe. Dort besteht die Aufgabenstellung für ein System darin, eine *Kostenfunktion* zu minimieren, die angibt, wie stark die Abweichung von einem gewünschten Verhalten ist. In der Wirtschaft gibt es Mechanismen und Richtlinien, die den *Nutzen* Einzelner, das *Wohlergehen* von Gruppen und den *Profit* von Unternehmen

optimiert.[10] In der Planungsforschung werden komplexe logistische und Fertigungsprobleme gelöst. Hierzu wird die erwartete *Summe der Belohnungen* im Laufe der Zeit maximiert. Und in der Statistik sind Lernalgorithmen darauf ausgelegt, eine erwartete *Verlustfunktion* zu minimieren, die die Kosten für Vorhersagefehler definiert.

Dieses allgemeine Schema – ich bezeichne es als *Standardmodell* – ist sehr verbreitet und extrem leistungsfähig. Doch unglücklicherweise *wünschen wir uns keine Maschinen, die auf diese Weise intelligent sind.*

Der Nachteil des Standardmodells wurde 1960 von Norbert Wiener, einem legendären MIT-Professor und führenden Mathematiker der Mitte des 20. Jahrhunderts, klar benannt. Wiener hatte gerade gesehen, wie das Dameprogramm von Arthur Samuel lernte, seinen Schöpfer zu besiegen. In der Folge schrieb er seinen visionären, aber kaum bekannten Artikel *Some Moral and Technical Consequences of Automation* (Ein wenig Moral und die technischen Folgen der Automatisierung).[11] Sein Hauptargument lautet:

> »Wenn wir zur Erlangung unserer Absichten einen maschinellen Agenten einsetzen, in dessen Handlungen wir nicht effektiv eingreifen können, [...] sollten wir uns absolut sicher sein, dass die der Maschine vorgegebene Absicht wirklich dem Ergebnis entspricht, das wir uns wünschen.«

»Die der Maschine vorgegebene Absicht« ist exakt das Ziel, das die Maschinen im Standardmodell zur Optimierung heranziehen. Wenn wir einer Maschine, die uns an Intelligenz übertrifft, die falsche Absicht vorgeben, setzt sie diese um und wir verlieren. Die Vorgänge in den sozialen Medien sind nur ein Vorgeschmack darauf. Hier optimieren relativ dumme Algorithmen in globalem Maßstab das falsche Ziel. In Kapitel 5 komme ich noch auf einige deutlich schlimmere Folgen zu sprechen.

All das dürfte niemanden wirklich überraschen. Seit Jahrtausenden kennen wir die Gefahren, die lauern, wenn wir genau das bekommen, was wir uns wünschen. In jedem Märchen, in dem drei Wünsche erfüllt werden, dient der dritte Wunsch dazu, die Folgen der ersten beiden rückgängig zu machen.

Es sieht so aus, als wäre unser Marsch in Richtung einer übermenschlichen Intelligenz unaufhaltsam. Kommen wir ans Ziel, könnte dies das Ende der Menschheit besiegeln. Doch noch ist nicht alles verlo-

ren. Wir müssen uns darüber klar werden, wo wir falsch abgebogen sind, und unseren Fehler beheben.

Gibt es Abhilfe?

Das Problem liegt in der grundlegenden Definition der künstlichen Intelligenz. Wir sagen, dass Maschinen insoweit intelligent sind, als ihre Handlungen generell darauf ausgerichtet sind, *ihre* Ziele zu erreichen. Aber wir haben keine zuverlässige Möglichkeit, sicherzustellen, dass *ihre* Ziele *unseren* Zielen entsprechen.

Vielleicht sollten wir vielmehr darauf bestehen, dass die Maschinen nicht *ihre* Absichten verfolgen, sondern *unsere* Absichten? Wenn wir eine solche Maschine jemals entwickeln können, ist sie nicht einfach nur *intelligent*, sondern auch *vorteilhaft* für die Menschen. Probieren wir es mit einer neuen Definition:

> Maschinen sind insoweit *vorteilhaft*, als *ihre* Handlungen darauf ausgerichtet sind, *unsere* Ziele zu erreichen.

Das hätte von Anfang an das Motto sein sollen.

Der schwierige Teil besteht natürlich darin, dass unsere Ziele und Absichten in uns liegen (und zwar in allen acht Milliarden von uns, die diesen Planeten in ihrer Vielfalt bevölkern) und nicht in den Maschinen. Dennoch ist es möglich, Maschinen zu konstruieren, die auf genau diese Weise zu unserem Wohl beitragen. Natürlich werden diese Maschinen unsere Absichten nicht ganz genau kennen. Wie sollten sie auch? Schließlich sind wir uns selbst oft genug nicht im Klaren darüber. Aber es wird sich zeigen, dass es sich bei dieser Unsicherheit um ein Feature handelt, nicht um einen Bug (das heißt, es ist an sich eine gute Sache, keine schlechte). Unsicherheit in puncto Ziele und Absichten bedeutet auch, dass Maschinen weiterhin dem Menschen folgen müssen: Sie werden um Erlaubnis fragen, sich korrigieren und auch abschalten lassen.

Indem wir die Voraussetzung streichen, dass Maschinen klare Ziele haben müssen, rütteln wir an den Grundfesten der künstlichen Intelligenz und reißen diese zum Teil ein. Das heißt, wir müssen den Überbau zum großen Teil neu errichten: die Ideen und Verfahren zur Umsetzung von KI. Das Ergebnis wird eine neue Beziehung zwischen Mensch und Maschine sein, die – so hoffe ich – dazu beiträgt, die nächsten paar Jahrzehnte erfolgreich zu meistern.

STICHWORTVERZEICHNIS

B

C

D

E

F

G

K

L

O

P

Q

R

S

W

Z