Die technologische Singularität

Murray Shanahan

Die technologische Singularität

Aus dem Englischen von Nadine Miller

Matthes & Seitz Berlin

Diese Gedanken kommen manchen Lesern
vielleicht fantastisch vor; dem Autor allerdings
erscheinen sie sehr real, von hoher Dringlichkeit
und auch außerhalb der Science-Fiction
bedenkenswert zu sein.

I. J. Good, *Speculations Concerning*
the First Ultraintelligent Machine (1965)

Echtes Motivproblem, bei einer KI. Eben kein Mensch.

William Gibson, *Neuromancer* (1984)

Inhaltsverzeichnis

Vorwort

Wie viele andere, die ihr berufliches Wirken der Erforschung der künstlichen Intelligenz gewidmet haben, bin ich als Kind von der Science-Fiction inspiriert worden. Die Heldin meiner jungen Jahre war keine reale Person, sondern Susan Calvin, die Wissenschaftlerin in Asimovs *I-Robot*-Geschichten (der Bücher, nicht der Verfilmung), eine Vorreiterin auf dem Gebiet der Roboterpsychologie. Wenn ich einmal groß wäre, wollte ich unbedingt so werden wie sie; heute, da ich (einigermaßen) erwachsen bin und im echten Leben den Titel eines Professors für Kognitive Robotik trage, ist meine Beziehung zur Science-Fiction allerdings etwas komplexer geworden. Zwar betrachte ich sie noch immer als Inspirationsquelle und als ein Medium, in dem wichtige philosophische Fragen erörtert werden können, doch die von ihr untersuchten Gegenstände verdienen eine eingehendere Behandlung. Das Hauptziel der Science-Fiction ist Unterhaltung, wenn auch auf eine intellektuell anregende Art und Weise. Es wäre jedoch verfehlt, sie als eine Denkanleitung zu betrachten.

Das vorliegende Buch ist daher weder als Werk der Science-Fiction noch als ein Beitrag zur sogenannten Futurologie zu verstehen. Sein Ziel ist nicht die Voraussage, sondern die Untersuchung einer Reihe von möglichen Zukunftsszenarien, ohne sich dabei auf das Eintreten eines bestimmten davon oder auf einen spezifischen zeitlichen Horizont dafür festzulegen. Tatsächlich ist es manchmal lohnend, auch höchst unwahrscheinliche oder abseitige Zukunftsszenarien näher zu betrachten. Dies gilt zum Beispiel dann, wenn es um eine besonders dystopische

Entwicklung geht. In diesem Fall könnten wir nämlich versucht sein, sehr genau darüber nachzudenken, wie wir die Wahrscheinlichkeit ihres Eintretens verringern können. Eine genauere Untersuchung unwahrscheinlicher oder abseitiger Szenarien lohnt sich außerdem auch dann, wenn sie interessante philosophische Fragen aufwirft, die uns etwa zum Nachdenken darüber nötigen, was wir als Spezies eigentlich wirklich wollen. Ganz gleich also, ob man glaubt oder nicht glaubt, dass wir schon bald eine künstliche Intelligenz auf menschlichem Niveau erschaffen werden oder dass die Singularität kurz bevorsteht – der Gedanke als solcher verdient es, ernsthaft in Erwägung gezogen zu werden.

Dies ist ein kurzes Buch über ein sehr großes Thema. Es kann daher höchstens als eine Einführung gelten, die viele wichtige Fragen nur anreißt. So werden hier zum Beispiel verschiedene Positionen in Bezug auf das Bewusstsein vorgestellt, zu denen es wohlbekannte Gegenpositionen gibt, die es ihrerseits verdienen, mit weiteren Gegenpositionen konfrontiert zu werden. Doch ein einführendes Werk muss solche Feinheiten übergehen. Sein Schwerpunkt liegt zudem eindeutig auf der Zukunft der künstlichen Intelligenz; einige wichtige damit zusammenhängende Themen wie Nanotechnologie und Biotechnologie werden jedoch nur am Rande angeschnitten. Das Buch soll einen neutralen Überblick über das konzeptuelle Territorium ermöglichen, und ich war bemüht, in strittigen Fällen die Positionen beider Seiten des Streits zu skizzieren. Allerdings wird es wohl, all meinen Bemühungen zum Trotz, unvermeidlich sein, dass einige meiner eigenen Ansichten durch den Schleier der Neutralität hindurchschimmern werden.

Ich möchte mich bei den vielen, vielen Menschen bedanken, die mit mir über die Jahrzehnte hinweg das Thema der künstlichen Intelligenz diskutiert haben, und zwar

nicht nur bei den Wissenschaftlern und Studenten, sondern auch bei denjenigen Vertretern der breiteren Öffentlichkeit, die meine Vorträge besucht haben. Gern würde ich jedem Einzelnen meinen Dank namentlich aussprechen, doch das ist natürlich nicht möglich. Deshalb werde ich meine explizite Danksagung ein paar Kollegen vorbehalten, deren Einfluss besonders in jüngster Zeit von besonderer Bedeutung gewesen ist. Ich danke also Stuart Armstrong, Nick Bostrom, Andrew Davison, Daniel Dewey, Randal Koene, Richard Newcombe, Owen Holland, Huw Price, Stuart Russell, Anders Sandberg und Jaan Tallinn. Alle, die ich hier zu erwähnen vergaß, bitte ich um Verzeihung. Abschließend möchte ich mich bei MIT Press bedanken, besonders bei Bob Prior, der mich ursprünglich dazu ermutigt hat, dieses Buch zu verfassen.

Murray Shanahan
North Norfolk und South Kensington, Oktober 2014

Einleitung

In den letzten Jahren ist die Vorstellung, dass sich die Menschheitsgeschichte aufgrund des immer schnelleren technologischen Fortschritts einer »Singularität« nähere, aus dem Reich der Science-Fiction in das der ernsthaften Diskussion gerückt. In der Physik bezeichnet »Singularität« einen bestimmten Punkt in Raum oder Zeit, etwa das Zentrum eines Schwarzen Lochs oder den Augenblick des Urknalls, an dem die Mathematik – und mit ihr unsere Fähigkeit zu begreifen – kollabiert. Analog dazu käme es in der menschlichen Geschichte zu einer Singularität, wenn ein exponentieller Fortschritt in der Technologie derart dramatische Veränderungen herbeiführen würde, dass die menschliche Existenz, wie wir sie heute verstehen, an ein Ende käme.[1] Die Institutionen, die wir für selbstverständlich halten – die Wirtschaft, die Regierung, das Rechtssystem und der Staat –, würden in ihrer jetzigen Form nicht überleben, die fundamentalsten menschlichen Werte – die Unantastbarkeit des Lebens, das Streben nach Glück, die Entscheidungsfreiheit – würden verdrängt werden, ja unsere ganze Auffassung davon, was es heißt, ein Mensch zu sein – nämlich ein Individuum zu sein, das lebendig, mit Bewusstsein ausgestattet und Teil einer sozialen Ordnung ist –, wäre radikal infrage gestellt, und das nicht etwa im Modus einer distanzierten philosophischen Betrachtung, sondern durch die Wucht der Umstände, ganz unmittelbar und real.

Welcher technologische Fortschritt könnte nun eine solche Umwälzung auslösen? In diesem Buch werden wir die Hypothese untersuchen, dass eine technologische Singularität dieser Art durch signifikante Fortschritte auf

einem von zwei miteinander zusammenhängenden Gebieten (oder auf beiden) herbeigeführt werden könnte, nämlich dem der KI-Forschung und dem der Neurotechnologie. Wir wissen bereits, wie wir am Stoff des Lebens, den Genen und der DNA, herumbasteln können, und die Auswirkungen der Biotechnologie sind für sich genommen schon gewaltig; wenn wir aber erst einmal gelernt haben, den »Stoff des Geistes« zu manipulieren, werden die möglichen Konsequenzen alles Vorangegangene in den Schatten stellen.

Der Intellekt ist heute in einem wichtigen Sinne erstarrt, was sowohl den Umfang als auch das Tempo des technologischen Fortschritts begrenzt. Natürlich wächst der menschliche Wissensschatz seit Jahrtausenden stetig an, und parallel dazu wächst dank der Erfindung der Schrift, des Buchdrucks und des Internets auch unsere Fähigkeit, dieses Wissen zu verbreiten. Dennoch ist das Organ, das Wissen produziert, nämlich das Gehirn des Homo sapiens, während dieser ganzen Zeit im Wesentlichen unverändert geblieben, und seine kognitiven Fähigkeiten sind nach wie vor unübertroffen.

Das wird sich allerdings ändern, wenn die Forschung auf dem Gebiet der künstlichen Intelligenz und der Neurotechnologie hält, was sie verspricht. Wenn der Intellekt nämlich nicht mehr nur Produzent der Technologie ist, sondern auch selbst zu ihrem Produkt wird, kann dies eine Feedbackschleife mit unabsehbaren und potenziell explosiven Konsequenzen zur Folge haben. Denn wenn das hergestellte Ding die Intelligenz selbst ist, also genau jene Entität, die diese Herstellung durchführt, dann kann sie sich anschicken, Verbesserungen an sich selbst vorzunehmen. Und der Singularitätshypothese zufolge ist der gewöhnliche Mensch denn auch bald aus dem Spiel, indem er entweder von KI-Maschinen oder von einer kognitiv

verbesserten biologischen Intelligenz überholt wird und nicht mehr mithalten kann.

Verdient es die Singularitätshypothese, dass wir sie ernst nehmen, oder ist sie nur eine mit viel Fantasie ersonnene Fiktion? Ein Argument dafür, sie ernst zu nehmen, gründet auf dem von Ray Kurzweil sogenannten Gesetz vom steigenden Ertragszuwachs [*law of accelerating returns*]: Ein technologischer Bereich untersteht diesem Gesetz, wenn das Tempo, mit dem die Technologie sich verbessert, sich proportional zu ihrer Qualität verhält. Mit anderen Worten, je besser die Technologie ist, umso schneller wird sie noch besser, was im Laufe der Zeit zu einer exponentiellen Verbesserung führt.

Ein prominentes Beispiel für dieses Phänomen ist das Moore'sche Gesetz, wonach sich die Anzahl der Transistoren, die auf einem einzigen Chip verfertigt werden können, etwa alle 18 Monate verdoppelt.[2] Es ist bemerkenswert, dass es der Halbleiterindustrie tatsächlich gelungen ist, dem Moore'schen Gesetz mehrere Jahrzehnte lang zu entsprechen. Andere Kennzahlen zur Bestimmung des Fortschritts in der Informationstechnologie, etwa die CPU-Taktfrequenz oder die Netzwerkbandbreite, haben sich ähnlich exponentiell entwickelt. Die IT ist jedoch nicht das einzige Gebiet, auf dem wir einen sich beschleunigenden Fortschritt beobachten können. In der Medizin etwa sind die Kosten für die DNA-Sequenzierung exponentiell gesunken, während ihre Geschwindigkeit exponentiell zunimmt, und die Hirnscantechnologie hat eine exponentielle Erhöhung der Bildauflösung zu verzeichnen.[3]

Auf einer historischen Zeitachse betrachtet präsentieren sich diese Trends zur Beschleunigung im Zusammenhang mit einer Reihe von technologischen Meilensteinen, die in immer kürzeren Abständen erreicht werden: Acker-

bau, Buchdruck, elektrische Energie, der Computer. Vor einem noch längeren, evolutionären Zeithorizont gesehen ging dieser Abfolge von technologischen Entwicklungen jedoch selbst schon eine Reihe evolutionärer Meilensteine voraus, die ebenfalls in immer kürzeren Abständen entstanden waren: Eukaryoten, Wirbeltiere, Primaten, der Homo sapiens. Angesichts dieser Tatsachen sind manche Experten der Meinung, dass die Entwicklung der menschlichen Gattung auf einer drastisch ansteigenden Komplexitätskurve voranschreitet, die bis in die fernste Vergangenheit zurückreicht. Doch wie dem auch sei, wir müssen nur denjenigen Abschnitt der Kurve ein wenig in die Zukunft weiterdenken, auf dem die Technologie angesiedelt ist, um an einen entscheidenden Kipppunkt zu gelangen, den Punkt nämlich, an dem menschliche Technologie den normalen Menschen in technologischer Hinsicht obsolet werden lässt.[4]

Natürlich erreicht jeder exponentielle technologische Trend irgendwann ein Plateau, einfach aufgrund der Gesetze der Physik, und es gibt zahllose ökonomische, politische oder wissenschaftliche Gründe, weshalb ein exponentiell verlaufender Trend ins Stocken geraten könnte, bevor er an sein theoretisches Limit gestoßen ist. Aber nehmen wir einmal an, dass die für die KI-Forschung und die Neurotechnologie relevantesten technologischen Trends ihre beschleunigte Dynamik beibehalten und uns die Fähigkeit verleihen, den »Stoff des Geistes« technisch zu erschaffen und die eigentliche Maschinerie der Intelligenz damit zu synthetisieren und zu manipulieren. An diesem Punkt unterläge die Intelligenz selbst, ob künstlich oder menschlich, dem Gesetz vom steigenden Ertragszuwachs, und um von dort aus zur technologischen Singularität zu gelangen, braucht es dann nur noch ein wenig Vertrauen in den Prozess.

Einige Autoren prophezeien voller Zuversicht, dass sich diese Zäsur Mitte des 21. Jahrhunderts ereignen wird. Doch auch abgesehen von der ohnehin unzuverlässigen Wahrsagerei gibt es gute Gründe, die Idee der Singularität ernsthaft zu durchdenken. Erstens ist von einem intellektuellen Standpunkt her betrachtet das Konzept als solches bereits hochinteressant, ganz unabhängig davon, ob oder wann sie jemals eintreten wird. Zweitens verlangt ihre bloße Möglichkeit – wie entfernt sie auch zu sein scheint – schon aus rein pragmatischen und gänzlich rationalen Gründen bereits heute nach einer Untersuchung. Auch wenn die Argumente der Futuristen nämlich nicht schlüssig sein sollten, es genügt schon, wenn wir dem vorhergesagten Ereignis auch nur die geringste Eintrittswahrscheinlichkeit zusprechen, damit es unsere gesamte, ungeteilte Aufmerksamkeit beanspruchen darf. Denn würde eine technologische Singularität tatsächlich eintreten, dann hätte dies für die Menschheit erdrutschartige Folgen.

Welches sind diese potenziell erdrutschartigen Folgen? Was für eine Welt, was für ein Universum entstünde, wenn sich eine technologische Singularität tatsächlich einstellte? Sollten wir ihr Eintreten fürchten oder es begrüßen? Was, wenn überhaupt, können wir heute oder in naher Zukunft tun, um den bestmöglichen Ausgang der ganzen Sache zu gewährleisten? Dies sind die wichtigsten der Fragen, die auf den folgenden Seiten behandelt werden. Diese Fragen sind zwar groß, doch die Aussicht auf die Singularität, ja sogar ihr bloßer Gedanke verspricht, uralte und vielleicht sogar noch größere philosophische Fragen in ein neues Licht zu rücken: Was ist der Kern unseres Menschseins? Welches sind unsere grundlegendsten Werte? Wie sollten wir leben, und worauf sind wir dabei bereit zu verzichten? Denn die Möglichkeit einer technologischen

Singularität stellt sowohl ein existenzielles Risiko als auch eine existenzielle Chance dar.

Ein existenzielles Risiko ist sie, weil sie für das schiere Überleben der menschlichen Gattung potenziell bedrohlich ist. Das klingt vielleicht übertrieben, aber die heute neu entwickelten Technologien verfügen über ein nie zuvor gesehenes Potenzial. So fällt es zum Beispiel nicht schwer, sich vorzustellen, dass ein hochgradig ansteckendes und arzneimittelresistentes Virus gentechnisch erzeugt werden könnte, das tödlich genug wäre, um eine solche Katastrophe herbeizuführen. Zwar würde nur ein Verrückter so etwas mit voller Absicht herstellen, aber möglicherweise braucht es nur etwas Leichtsinn, um ein Virus zu erschaffen, das in der Lage wäre, zu einem solchen Monster zu mutieren. Eine fortgeschrittene KI könnte nun aus analogen, aber weitaus subtileren Gründen eine existenzielle Gefahr darstellen. Wir werden zu gegebener Zeit auf diese zu sprechen kommen. Für den Moment genügt es uns festzustellen, dass es absolut vernünftig ist, über die Möglichkeit nachzudenken, dass irgendein Konzern, eine Regierung, eine Organisation oder sogar eine Einzelperson in der Zukunft eine sich exponentiell selbstverbessernde, ressourcenhungrige KI erschafft und dann die Kontrolle über sie verliert.

Aus einem etwas optimistischeren Blickwinkel betrachtet könnte eine technologische Singularität aber auch als eine existenzielle Chance im eher philosophischen Sinne des Wortes existenziell angesehen werden. Die Fähigkeit, den »Stoff des Geistes« technisch herzustellen, verschafft uns nämlich die Möglichkeit, unser biologisches Erbe zu transzendieren und dabei die mit ihm verbundenen Limitierungen zu überwinden. An erster Stelle dieser Limitierungen steht die Sterblichkeit. Der Körper eines Tieres ist ein empfindliches Gebilde, das anfällig ist

für Krankheiten, Verletzungen und Verfall, und das biologische Gehirn, an das das menschliche Bewusstsein (gegenwärtig noch) gebunden ist, ist einfach nur eines seiner Teile. Sollten wir aber einmal der Mittel habhaft werden, um Beschädigungen an ihm gleich welchen Schweregrads reparieren und das Gehirn schließlich von Grund auf (womöglich in einem nichtbiologischen Substrat) nachbauen zu können, stünde einer unbegrenzten Erweiterung des Bewusstseins nichts Grundsätzliches mehr im Wege.

Die Verlängerung des Lebens ist ein Aspekt eines Trends, der als Transhumanismus bezeichnet wird. Doch warum sollten wir uns mit dem menschlichen Leben in der Form, in der wir es kennen, zufriedengeben? Wenn wir das Gehirn nachbauen können, wieso sollten wir dann nicht auch in der Lage sein, es umzugestalten und zu verbessern? (Die gleiche Frage könnte man übrigens auch mit Blick auf den menschlichen Körper stellen, aber unser Thema hier ist der Intellekt.) Konservative Optimierungen der Gedächtnisleistung, der Lernfähigkeit und der Aufmerksamkeit lassen sich bereits jetzt mit pharmazeutischen Mitteln erzielen; das Vermögen, das Gehirn von Grund auf umzubauen, deutet allerdings auf Möglichkeiten für radikalere Formen der kognitiven Verbesserung und Umstrukturierung hin. Was könnten oder sollten wir mit solchen transformativen Kräften anfangen? Nun, zumindest würden sie, wie manchmal behauptet wird, die existenzielle Gefährdung durch superintelligente Maschinen reduzieren. Sie würden es uns also ermöglichen, mit der Entwicklung Schritt zu halten, obwohl wir uns im Laufe des Prozesses vielleicht bis zur Unkenntlichkeit verändern würden.

Um den umfassendsten – und provokantesten – Sinn zu erfassen, in dem eine technologische Singularität eine existenzielle Chance darstellen könnte, müssen wir uns gänzlich von der menschlichen Perspektive verabschieden

und einen eher kosmologischen Standpunkt einnehmen. Es ist sicherlich die Krönung des anthropozentrischen Denkens anzunehmen, dass die Geschichte der Materie in dieser unserer Ecke des Universums in der menschlichen Gesellschaft und den unzähligen darin eingebetteten lebendigen Gehirnen gipfelt, so wunderbar diese auch sein mögen. Vielleicht steht der Materie auf der Komplexitätsskala noch ein langer Weg nach oben bevor. Vielleicht gibt es Formen des Bewusstseins, die erst noch entstehen werden und die dem unsrigen in gewissem Sinne überlegen sind. Sollten wir vor dieser Aussicht zurückschrecken oder sie bejubeln? Könnten wir einen solchen Gedanken überhaupt gänzlich begreifen? Diese Fragen verdienen es, erörtert zu werden, ganz gleich, ob die Singularität nahe ist oder nicht – zumal der Versuch ihrer Beantwortung ein neues Licht auf uns selbst und auf unsere Stellung in der Ordnung der Dinge wirft.

Kapitel 1
Wege zur künstlichen Intelligenz

1.1 Allgemeine künstliche Intelligenz

1950 veröffentlichte Alan Turing, der während des Zweiten Weltkriegs als Codebrecher tätig war und als Pionier der Informatik gilt, in der Zeitschrift *Mind* einen Aufsatz mit dem Titel »Computing Machinery and Intelligence« [»Kann eine Maschine denken?«].[5] Dies war die erste ernsthafte, wissenschaftliche Abhandlung über das Konzept der künstlichen Intelligenz. Turing sagte voraus, dass man im Jahr 2000 »widerspruchslos von denkenden Maschinen reden kann«, und stellte sich vor, dass Maschinen zu jenem Zeitpunkt die Prüfung würden bestehen können, die wir heute als den Turing-Test kennen.

Der Turing-Test ist eine Art Spiel. Zwei »Spieler«, ein Mensch und eine Maschine, kommunizieren dabei mit einem Dritten, dem »Schiedsrichter«, vermittels Tastatur und Bildschirm. Der Schiedsrichter führt nacheinander mit jedem der Spieler ein Gespräch und versucht zu erraten, welcher von beiden der Mensch und welcher die Maschine ist. Die Aufgabe der Maschine ist es, den Schiedsrichter davon zu überzeugen, dass sie ein Mensch ist – eine Leistung, die, wie es heißt, gewiss eine Intelligenz auf menschlichem Niveau erfordert. Kann der Schiedsrichter den Menschen nicht von der Maschine unterscheiden, dann hat sie den Test bestanden. Und als Turing dies im Jahre 1950 schrieb, antizipierte er eine Welt, in der Maschinen, die seinen Test bestehen könnten, etwas Alltägliches sein würden, »Denkmaschinen«

im Haushalt und am Arbeitsplatz also völlig normal wären.

Turings Prognose zum Trotz gab es bis zum Jahr 2000 allerdings weder eine KI auf menschlichem Niveau noch Anzeichen dafür, dass sie in absehbarer Zeit zu erwarten wäre. Keiner Maschine gelang es auch nur annähernd, den Turing-Test zu bestehen. Dennoch hatte man unlängst einen wichtigen Meilenstein in Sachen künstliche Intelligenz erreicht. Denn im Jahr 1997 hatte Deep Blue, ein IBM-Computer, den damaligen Schachweltmeister Garry Kasparow besiegt. Anders als bei früheren Schachprogrammen, die er geschlagen hatte und die ihm berechenbar und mechanisch erschienen waren, soll Kasparow über Deep Blue gesagt haben, er habe im Spiel eine »fremde Intelligenz« auf der anderen Seite des Schachbretts wahrgenommen.[6]

Es ist aufschlussreich, kurz innezuhalten und diesen Augenblick in der Geschichte der KI zu reflektieren. Denn auf diesem Gebiet war etwas erreicht worden, das ein halbes Jahrhundert zuvor vielleicht als sein krönender Abschluss gegolten hätte: Der Mensch war von einer Maschine überflügelt worden. Natürlich fährt auch ein Auto schneller, als der schnellste menschliche Sprinter laufen kann, und ein Baukran bewegt weit mehr Kilogramm in die Höhe als ein Weltmeister im Gewichtheben. Es sind aber seine intellektuellen Fähigkeiten, die den Menschen von der übrigen Tierwelt abheben, und das Schachspiel ist nun mal ein ausgesprochen intellektuelles Unterfangen.

Das Computerschach war also geknackt, und doch schien es so, als wären wir einer KI auf menschlichem Niveau in keiner Weise nähergekommen als zu Turings Zeit. Wie konnte das sein? Das Problem mit Deep Blue war seine Spezialisierung. Der Computer konnte nichts anderes als Schach spielen. Man vergleiche ihn mit einem