


Szenarien
einer
kommenden
Revolution

Nick
Bostrom
Super
intelligenz



Suhrkamp

Was geschieht, wenn es uns eines Tages gelingt, eine Maschine zu entwickeln, die die menschliche Intelligenz auf so gut wie allen Gebieten übertrifft? Klar ist: Eine solche Superintelligenz wäre enorm mächtig und würde uns vor riesige Kontroll- und Steuerungsprobleme stellen. Mehr noch: Vermutlich würde die Zukunft der menschlichen Spezies in den Händen dieser Superintelligenz liegen, so wie heute die Zukunft der Gorillas von uns abhängt.

Nick Bostrom nimmt uns mit auf eine faszinierende Reise in die Welt der Orakel und Genies, der Superrechner und Gehirnsimulationen, aber vor allem in die Labore dieser Welt, in denen derzeit fieberhaft an der Entwicklung einer künstlichen Intelligenz gearbeitet wird. Er skizziert mögliche Szenarien, wie die Geburt der Superintelligenz vonstattengehen könnte, und widmet sich ausführlich den Folgen dieser Revolution.

Sie werden global sein und unser wirtschaftliches, soziales und politisches Leben tiefgreifend verändern. Wir müssen handeln, und zwar kollektiv, bevor der Geist aus der Flasche gelassen ist – also jetzt! Das ist die eminent politische Botschaft dieses so spannenden wie wichtigen Buches.

Nick Bostrom, geboren 1973, ist Professor für Philosophie am St. Cross College der Universität von Oxford und Direktor sowohl des Future of Humanity Institute als auch des Programme for the Impact of Future Technology, die beide Teil der Oxford Martin School sind. 2009 wurde er für seine Arbeit mit dem prestigeträchtigen Eugene R. Gannon Award for the Continued Pursuit of Human Advancement ausgezeichnet und war auf der *100 Top Global Thinkers List* von *Foreign Policy*.

Nick Bostrom

Superintelligenz

Szenarien einer kommenden Revolution

Aus dem Englischen von
Jan-Erik Strasser

Suhrkamp

Titel der Originalausgabe: *Superintelligence. Paths, Dangers, Strategies*

First Edition was originally published in English in 2014.

This translation is published by arrangement with Oxford University Press.

Erstmals erschienen 2014 bei Oxford University Press. Die Übersetzung erscheint mit freundlicher Genehmigung von Oxford University Press.

Copyright © Nick Bostrom 2014

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation

in der Deutschen Nationalbibliografie;

detaillierte bibliografische Daten sind im Internet

über <http://dnb.d-nb.de> abrufbar.

4. Auflage 2020

Wissenschaftliche Sonderausgabe

Erste Auflage 2016

© der deutschen Ausgabe Suhrkamp Verlag Berlin 2014

© Nick Bostrom 2014

© dieser Ausgabe Suhrkamp Verlag Berlin 2016

Alle Rechte vorbehalten, insbesondere das der Übersetzung, des öffentlichen Vortrags sowie der Übertragung durch Rundfunk und Fernsehen, auch einzelner Teile. Kein Teil des Werkes darf in irgendeiner Form (durch Fotografie, Mikrofilm oder andere Verfahren) ohne schriftliche Genehmigung des Verlages reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden.

Einbandgestaltung: Hermann Michels und Regina Göllner

Satz: Satz-Offizin Hümmer GmbH, Waldbüttelbrunn

Druck: Kösel, Krugzell

Printed in Germany

ISBN 978-3-518-58684-6

Inhalt

Die unvollendete Fabel von den Spatzen	7
Vorwort	9
1. Vergangene Entwicklungen und gegenwärtige Möglichkeiten .	13
2. Wege zur Superintelligenz	41
3. Formen der Superintelligenz	80
4. Die Kinetik einer Intelligenzexplosion	93
5. Der entscheidende strategische Vorteil	115
6. Kognitive Superkräfte	131
7. Der superintelligente Wille	149
8. Sind wir dem Untergang geweiht?	164
9. Das Kontrollproblem	181
10. Orakel, Flaschengeister, Souveräne und Werkzeuge	206
11. Multipolare Szenarien	224
12. Der Erwerb von Werten	260
13. Die Wahl der Auswahlkriterien	292
14. Das strategische Gesamtbild	320
15. Die heiße Phase	358
Anmerkungen	367
Danksagung	439
Verzeichnis der Abbildungen, Tabellen und Kästen	440
Literaturverzeichnis	442
Register	469
Ausführliches Inhaltsverzeichnis	477

Die unvollendete Fabel von den Spatzen

Es war die Zeit des Nestbaus, aber nach tagelanger harter Arbeit saßen die Spatzen in der Abenddämmerung beisammen, um auszuruhen und miteinander zu zwitschern.

»Wir sind alle so klein und schwach. Stellt euch vor, wie angenehm es wäre, wenn wir eine Eule hätten, die uns beim Nestbau helfen würde!«

»Genau!«, stimmte ein anderer Spatz ein. »Sie könnte bei Jung und Alt nach dem Rechten sehen.«

»Und uns Ratschläge geben und vor der Nachbarskatze warnen«, meinte ein dritter.

Darauf sprach Pastus, der Schwarmälteste: »Schicken wir unsere Späher in alle Himmelsrichtungen aus, um ein verwaistes Eulenküken oder ein Ei zu finden. Auch ein Krähenjunges oder ein kleines Wiesel könnten von Nutzen sein. Das ist vielleicht das Beste, was uns je passiert ist – zumindest seit der Eröffnung des Pavillons der nie enden wollenden Körner.«

Die ganze Schar war begeistert, und die Spatzen allüberall begannen aus vollster Kehle zu trällern.

Nur Scronkfinkle, ein einäugiger Spatz von mürrischem Gemüt, war von der Klugheit des Vorhabens nicht überzeugt. Er sprach: »Das wird unser Verderben sein. Sollten wir nicht erst bedenken, wie eine Eule sich zähmen und bändigen lässt, bevor wir sie in unsere Mitte bringen?«

Pastus erwiderte: »Eine Eule zu zähmen scheint mir ein höchst schwieriges Unterfangen zu sein. Es wird schon schwer genug werden, ein Ei zu finden. Fangen wir also damit an. Nachdem wir die Eule großgezogen haben, können wir uns der nächsten Herausforderung stellen.«

»Der Plan hat einen Haken!«, piepste Scronkfinkle, aber sein Protest ging im Tumult des auffliegenden Schwarms unter, der sich daranmachte, Pastus' Vorhaben in die Tat umzusetzen.

Nur zwei oder drei Spatzen blieben mit Scronkfinkle zurück. Gemeinsam versuchten sie herauszufinden, wie Eulen gezähmt oder gebändigt werden könnten. Schon bald wurde ihnen klar, dass Pastus recht gehabt hatte: Es war in der Tat ein höchst schwieriges Unterfangen, zu-

mal ihnen eine Eule zum Üben fehlte. Nichtsdestotrotz fuhren sie fort, so gut sie konnten, und in der ständigen Furcht, die anderen Spatzen könnten mit einem Ei zurückkehren, bevor eine Lösung für das Kontrollproblem gefunden wäre.

Niemand weiß, wie die Geschichte ausgeht, aber der Autor widmet dieses Buch Scronkfinkle und seinen Anhängern.

Vorwort

In Ihrem Schädel steckt etwas, das diesen Satz liest. Dieses Ding, das menschliche Gehirn, hat einige Fähigkeiten, die den Gehirnen anderer Tiere fehlen, und diesen besonderen Fähigkeiten verdanken wir unsere dominante Stellung auf der Erde. Andere Tiere haben stärkere Muskeln oder schärfere Krallen, doch wir haben die schlaueren Gehirne. Dieser bescheidene Vorteil an allgemeiner Intelligenz hat uns schließlich Sprache, Technologie und komplexe soziale Organisationen entwickeln lassen, und er wuchs im Lauf der Zeit, da jede Generation auf den Leistungen der vorigen aufbauen konnte.

Falls wir eines Tages künstliche Gehirne bauen, die das menschliche an allgemeiner Intelligenz übertreffen, dann könnte diese neue Art von Superintelligenz überaus mächtig werden. Und genau wie das Schicksal der Gorillas heute stärker von uns Menschen abhängt als von den Gorillas selbst, so hinge das Schicksal unserer Spezies von den Handlungen dieser maschinellen Superintelligenz ab.

Wir haben allerdings einen Vorteil: Wir sind diejenigen, die das Ding bauen. Im Prinzip könnten wir eine Art von Superintelligenz erschaffen, die menschliche Werte achtet, und wir hätten sicherlich gute Gründe, genau das zu tun. In der Praxis jedoch sieht das Kontrollproblem (wie können wir kontrollieren, was die Superintelligenz tun würde?) ziemlich schwierig aus. Außerdem scheint es, als hätten wir nur einen Schuss frei. Sobald eine unfreundliche Superintelligenz existiert, wird sie uns davon abhalten, sie zu ersetzen oder ihre Präferenzen zu ändern. Dann wäre unser Schicksal besiegelt.

In diesem Buch versuche ich zu verstehen, welche Herausforderung die Superintelligenz darstellt und wie wir darauf reagieren sollten. Dies ist die wahrscheinlich größte und beängstigendste Aufgabe, der die Menschheit je gegenüberstand – und egal, ob wir sie meistern oder an ihr scheitern: Es wird wohl auch die letzte sein.

Sie werden hier keine Argumente dafür finden, dass wir kurz vor einem großen Durchbruch in der Forschung zur künstlichen Intelligenz stehen oder dass sich auch nur einigermaßen genau vorhersagen lässt,

wann es so weit ist. Es sieht so aus, als ob es irgendwann in diesem Jahrhundert geschehen wird, aber sicher können wir uns dessen nicht sein.

Die ersten Kapitel zeigen mögliche Wege zur Superintelligenz auf und sagen etwas zum zeitlichen Ablauf, doch der größte Teil des Buches ist der Frage gewidmet, was danach geschieht. Wir untersuchen die Kinetik einer Intelligenzexplosion, die Formen und die Fähigkeiten einer Superintelligenz sowie die Strategien, die einem superintelligenten Akteur zur Verfügung stehen, sobald er einen entscheidenden Vorteil erlangt hat. Im Anschluss verlagern wir unseren Fokus auf das Kontrollproblem und fragen danach, was zu tun ist, damit wir das Resultat all dieser Entwicklungen überleben und es für uns günstig ausfällt. Gegen Ende des Buches treten wir schließlich einen Schritt zurück, betrachten das große Ganze, das sich aus unseren Untersuchungen ergeben hat, und machen einige Vorschläge dazu, was schon jetzt getan werden kann, um eine existentielle Katastrophe zu vermeiden.

Es war nicht leicht, dieses Buch zu schreiben. Ich hoffe, dass der nun freigeräumte Weg es anderen Forschern ermöglicht, das neue Territorium schneller und bequemer zu erreichen, sodass sie sich dort – frisch und ausgeruht – mit uns daranmachen können, die Grenzen unseres Verständnisses zu erweitern. (Und falls dieser Weg ein wenig holprig und kurvenreich ist, so hoffe ich, dass die Kritiker die ursprüngliche Unwirtlichkeit des Geländes im Nachhinein nicht unterschätzen!)

Es war nicht leicht, dieses Buch zu schreiben: Ich habe versucht, es lesbar zu machen, aber es ist mir wohl nicht ganz gelungen. Beim Schreiben hatte ich als Zielgruppe ein früheres Ich im Sinn, das am Lesen dieses Buches Gefallen gefunden hätte – und das könnte den Leserkreis ganz schön einschränken. Dennoch denke ich, dass der Inhalt vielen Menschen zugänglich ist, wenn sie ihn gründlich studieren und der Versuchung widerstehen, jeden neuen Gedanken augenblicklich mit dem nächstliegenden Klischee zu verwechseln. Leser ohne entsprechendes Vorwissen sollten sich nicht von den paar Happen Mathematik oder Fachvokabular abschrecken lassen, denn es ist immer möglich, den Hauptgedanken aus dem Kontext zu erschließen (und umgekehrt werden diejenigen Leser, die ans Eingemachte wollen, in den Anmerkungen fündig werden)¹.

Viele der Argumente in diesem Buch sind vermutlich falsch,² und wahrscheinlich habe ich auch Überlegungen von entscheidender Bedeu-

tung nicht berücksichtigt, womit einige oder alle meine Schlussfolgerungen ungültig wären. Es hat mich einige Mühe gekostet, im gesamten Text auf Nuancen und Grade der Ungewissheit hinzuweisen, indem ich ihn mit einer Unzahl von Wörtern wie »könnte«, »dürfte«, »müsste«, »vielleicht« oder »wahrscheinlich« gespickt habe. Jede Einschränkung dieser Art wurde sorgfältig und bewusst platziert, doch selbst diese Hinweise epistemischer Bescheidenheit reichen nicht aus – sie müssen um ein globales Eingeständnis der Ungewissheit und Fehlbarkeit ergänzt werden. Das ist keine falsche Bescheidenheit: Obwohl mein Buch wahrscheinlich schwere Fehler und Irreführungen enthält, stehen die in der Literatur vorgebrachten Alternativen meiner Meinung nach noch wesentlich schlechter da – einschließlich der üblichen Ansicht (der »Nullhypothese«), der zufolge wir die Aussicht auf eine Superintelligenz vorerst gefahrlos oder vernünftigerweise ignorieren können.

1. Vergangene Entwicklungen und gegenwärtige Möglichkeiten

Ganz zu Beginn werfen wir einen Blick zurück. Wenn man sich die historische Entwicklung im größtmöglichen Maßstab ansieht, so scheint sie sich als Abfolge unterschiedlicher und immer schnellerer Wachstumsmodi darzustellen. Aufgrund dieses Musters wurde angenommen, dass ein weiterer (noch schnellerer) Wachstumsmodus möglich sein könnte. Wir messen dieser Beobachtung allerdings nicht zu viel Gewicht bei: Dies ist kein Buch über »technologische Beschleunigung«, »exponentielles Wachstum« oder die diversen Vorstellungen, die manchmal unter der Rubrik der »Singularität« versammelt werden. Im Anschluss an diese Überlegungen betrachten wir die Geschichte der KI-Forschung, bevor wir uns einen Überblick über den aktuellen Stand dieser Disziplin verschaffen. Zuletzt sehen wir uns einige kürzlich durchgeführte Expertenbefragungen an und setzen uns mit unserer Unwissenheit bezüglich der Zeitpunkte zukünftiger Fortschritte auseinander.

Wachstumsmodi und Weltgeschichte

Es ist erst einige Millionen Jahre her, dass unsere Vorfahren sich durch die Baumkronen der afrikanischen Urwälder schlangen. Nach geologischen und sogar evolutionären Maßstäben ist der Homo sapiens schnell aus seinem letzten gemeinsamen Vorfahren mit den Menschenaffen hervorgegangen. Wir entwickelten die aufrechte Haltung, den opponierbaren Daumen und – am wichtigsten – einige relativ geringfügige Änderungen der Gehirngröße und neurologischen Organisation, die zu einem sprunghaften Anstieg der kognitiven Fähigkeiten führten. Als Folge davon kann der Mensch abstrakt denken, komplexe Gedanken ausdrücken und Informationen über die Generationen hinweg kulturell weit besser anhäufen als jede andere Spezies auf diesem Planeten.

Diese Fähigkeiten ließen uns immer effizientere Produktionsmetho-

den entwickeln, sodass es unseren Vorfahren schließlich gelang, den Regenwald und die Savanne weit hinter sich zu lassen. Insbesondere nach der Einführung der Landwirtschaft wuchs die Bevölkerungsdichte mit der Gesamtbevölkerung an. Mehr Menschen bedeuteten auch mehr Ideen; größere Dichten bedeuteten, dass Ideen sich leichter ausbreiten und manche Menschen sich spezialisieren konnten. Diese Entwicklungen erhöhten die *Wachstumsraten* der wirtschaftlichen Produktivität und der technologischen Leistungsfähigkeit. Spätere, mit der industriellen Revolution zusammenhängende Entwicklungen führten zu einem zweiten, vergleichbar dramatischen Anstieg der Wachstumsrate.

Solche Anstiege haben gewichtige Konsequenzen. Vor ein paar hunderttausend Jahren, in der frühen menschlichen (oder hominiden) Vorgeschichte, war das Wachstum so langsam, dass es etwa eine Million Jahre gedauert hätte, um die Produktionskapazitäten so weit zu erhöhen, dass das Existenzminimum einer weiteren Million Menschen gesichert gewesen wäre. Bis zum Jahr 5000 v. Chr. stieg die Wachstumsrate in der Folge der Erfindung der Landwirtschaft dann so weit an, dass das gleiche Wachstum in nur zwei Jahrhunderten erreicht wurde. Heute, nach der industriellen Revolution, dauert es noch 90 Minuten.¹

Auch die gegenwärtige Wachstumsrate wird also schon eindrucksvolle Ergebnisse liefern, wenn sie noch eine gewisse Zeit stabil bleibt. Wächst die Wirtschaft weiterhin so wie in den letzten 50 Jahren, dann wird die Welt im Jahr 2050 rund 4,8-mal reicher und im Jahr 2100 etwa 34-mal reicher sein als heute.²

Doch die Aussicht auf ein stabiles exponentielles Wachstum verblasst im Vergleich zu einem erneuten sprunghaften *Anstieg der Wachstumsrate*, der mit denjenigen der landwirtschaftlichen oder der industriellen Revolution vergleichbar wäre. Auf der Grundlage historischer Wirtschafts- und Bevölkerungsdaten hat der Ökonom Robin Hanson die Verdopplungszeit des Weltwirtschaftswachstums geschätzt. Für Gemeinschaften von Jägern und Sammlern im Pleistozän kommt er auf 224 000 Jahre, für Agrargesellschaften auf 909 Jahre und für Industriegesellschaften auf 6,3 Jahre.⁴ (Hansons Modell zufolge herrscht in der gegenwärtigen Epoche eine Mischung aus landwirtschaftlichen und industriellen Wachstumsmodi vor – die Weltwirtschaft als ganze verdoppelt sich noch nicht alle 6,3 Jahre.) Bei einem weiteren Übergang zu

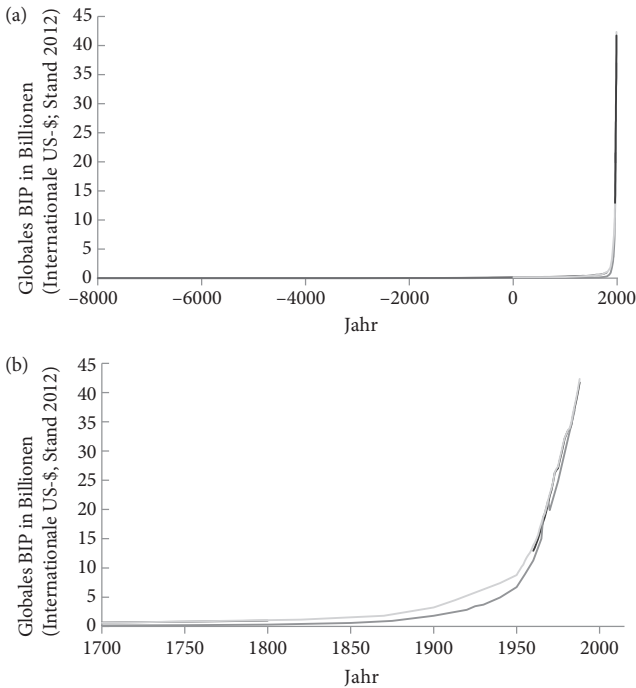


Abb. 1: Langzeitbetrachtung des globalen BIP. Auf einer linearen Skala aufgetragen, ähnelt die Geschichte der Weltwirtschaft einer flachen Linie, die sich an die x-Achse schmiegt, bis sie plötzlich senkrecht nach oben schießt. (a) Selbst wenn wir uns nur auf die letzten 10 000 Jahre beschränken, bleibt das Muster im Wesentlichen ein 90-Grad-Winkel. (b) Erst in den letzten rund 100 Jahren steigt die Kurve sichtbar an. (Die verschiedenen Linien des Graphen entsprechen verschiedenen Datensätzen, die leicht unterschiedliche Schätzwerte liefern.³)

einem anderen Wachstumsmodus ähnlichen Ausmaßes wäre eine sich etwa alle zwei Wochen verdoppelnde Weltwirtschaft die Folge.

So etwas erscheint derzeit unvorstellbar, aber in früheren Zeiten mag die Annahme ebenso absurd gewesen sein, die Weltwirtschaft könne sich irgendwann binnen eines einzigen Menschenlebens mehrmals verdoppeln – und doch halten wir genau diese außergewöhnliche Situation heute für völlig normal.

Der Gedanke einer kommenden technologischen Singularität ist mittlerweile weit verbreitet, wofür sowohl Vernor Vinges bahnbrechender Aufsatz als auch die Schriften von Ray Kurzweil und anderen gesorgt haben.⁵ Der Begriff »Singularität« wurde jedoch in vielen unterschied-

lichen Bedeutungen ge- und missbraucht und ist mittlerweile von einer unheiligen (wenngleich fast millenaristischen) Aura techno-utopischer Konnotationen umgeben.⁶ Da das meiste davon für unsere Argumentation ohne Belang ist, können wir Klarheit schaffen, indem wir auf das Wort »Singularität« verzichten und stattdessen präzisere Begriffe nutzen.

Die mit der Singularität zusammenhängende Idee, die uns hier interessiert, ist die Möglichkeit einer *Intelligenzexplosion*, insbesondere die Aussicht auf eine maschinelle Superintelligenz. Es mag Leute geben, die durch Wachstumsdiagramme (wie die in *Abbildung 1* präsentierten) davon überzeugt wurden, dass ein weiterer Wachstumsmodus vor der Tür steht, der mit den landwirtschaftlichen oder industriellen Revolutionen vergleichbar ist. Diese Leute könnten daraus schließen, dass ein Szenario, in dem die Verdopplungszeit der Weltwirtschaft nur wenige Wochen beträgt, die Erschaffung von Intelligenzen voraussetzt, die viel schneller und effizienter sind als die uns vertrauten biologischen. Um die Möglichkeit einer solchen Revolution ernst zu nehmen, müssen wir jedoch keine Punkte auf Kurven auftragen oder Wirtschaftswachstumsdaten extrapolieren. Wie wir sehen werden, gibt es dafür stärkere Gründe.

Große Erwartungen

Mit Maschinen, die dem Menschen an allgemeiner Intelligenz gleichkommen – die also über gesunden Menschenverstand ebenso verfügen wie über die Fähigkeit zu lernen, zu schlussfolgern und zu planen, um komplexe Herausforderungen in einer Vielzahl von natürlichen und abstrakten Bereichen zu meistern –, wurde seit der Erfindung des Computers in den 1940er Jahren gerechnet. Zu jener Zeit wurde das Erscheinen solcher Maschinen oft etwa 20 Jahre in die Zukunft verlegt.⁷ Seitdem hat sich das erwartete Ankunftsdatum mit einer Geschwindigkeit von einem Jahr pro Jahr nach hinten verschoben, sodass manche Futuristen, die sich mit der Möglichkeit der künstlichen allgemeinen Intelligenz befassen, auch heute noch glauben, dass es binnen zwei Jahrzehnten intelligente Maschinen geben wird.⁸

Zwei Dekaden sind eine ideale Zeitspanne für Propheten eines radi-

kalen Wandels: kurz genug, um Aufmerksamkeit zu erregen und relevant zu sein, aber lang genug, um zu plausibilisieren, dass bis dahin eine Reihe von Durchbrüchen stattgefunden haben könnte, die derzeit nur zu erahnen sind. Zum Vergleich: Die meisten Technologien, die in fünf oder zehn Jahren einen großen Einfluss auf die Welt haben werden, sind bereits in begrenztem Einsatz, und diejenigen, die die Welt binnen 15 Jahren revolutionieren, existieren wahrscheinlich schon als Prototypen. Zwanzig Jahre dürften auch in der Nähe des Karriereendes eines typischen Prognostikers liegen, was das Risiko mindert, mit einer kühnen Vorhersage den eigenen Ruf zu ruinieren.

Aus der Tatsache, dass viele Prognosen in der Vergangenheit zu vorschleunigt waren, folgt allerdings nicht, dass eine künstliche Intelligenz (KI) unmöglich ist oder niemals entwickelt werden wird.⁹ Der Hauptgrund für die Verzögerung ist, dass die technischen Probleme bei der Konstruktion intelligenter Maschinen größer sind, als die frühen Pioniere dachten. Dies aber lässt das konkrete Ausmaß der Probleme ebenso offen wie die Frage, wie weit wir jetzt von deren Überwindung entfernt sind. Manchmal hat ein Problem, das zunächst hoffnungslos kompliziert erscheint, eine überraschend einfache Lösung (zugegebenermaßen kommt das Umgekehrte vermutlich häufiger vor).

Im nächsten Kapitel werden wir verschiedene Wege betrachten, die zu einer maschinellen Intelligenz auf menschlichem Niveau führen könnten. Aber schon an dieser Stelle sei daran erinnert, dass diese nicht das Ziel der Reise darstellt, egal, wie viele Zwischenstopps wir auf dem Weg dorthin einlegen müssen. Bereits der nächste Halt danach, nur eine kurze Strecke entfernt, heißt *übermenschliche* maschinelle Intelligenz. Der Zug wird in Menschendorf nicht anhalten oder auch nur abbremsen, sondern wahrscheinlich einfach durchrasen.

Der Mathematiker I. J. Good, der im Zweiten Weltkrieg Chefstatistiker der Gruppe um Alan Turing war, die die deutschen Geheimcodes entschlüsselte, dürfte die zentralen Aspekte dieses Szenarios als Erster formuliert haben. In einer häufig zitierten Passage aus dem Jahr 1965 schrieb er:

Eine ultraintelligente Maschine sei definiert als eine Maschine, die alle geistigen Anstrengungen jedes noch so schlaun Menschen bei weitem übertreffen kann. Da die Konstruktion von Maschinen solch eine geistige Anstrengung ist, könnte eine ultraintelligente Maschine noch bessere Maschinen

konstruieren; zweifellos würde es dann zu einer »Intelligenzexplosion« kommen, und die menschliche Intelligenz würde weit dahinter zurückbleiben. Die erste ultraintelligente Maschine ist also die letzte Erfindung, die der Mensch je machen muss, vorausgesetzt, die Maschine ist fügsam genug, um uns zu sagen, wie man sie unter Kontrolle hält.¹⁰

Heute scheint es offensichtlich zu sein, dass große existentielle Risiken mit einer solchen Intelligenzexplosion verbunden wären und das Thema daher mit größtem Ernst zu behandeln ist, selbst wenn klar wäre (was nicht der Fall ist), dass nur eine relativ kleine Wahrscheinlichkeit dafür besteht. Die meisten Pioniere der künstlichen Intelligenz aber zogen die Möglichkeit einer übermenschlichen KI nicht in Betracht, obwohl sie von einer unmittelbar bevorstehenden KI menschlichen Niveaus überzeugt waren. Es ist, als ob ihr Mut zum Mutmaßen vom Er-sinnen der radikalen Möglichkeit intelligenter Maschinen so erschöpft worden wäre, dass sie sich das Korollar – superintelligente Maschinen – nicht mehr ausmalen konnten.

Auch die Möglichkeit von Risiken ließen sie nicht zu,¹¹ ja, sie taten nicht einmal so, als ob irgendwelche Sicherheitsbedenken oder moralische Skrupel bei der Schaffung von künstlichen Intelligenzen und potentiellen Computerdiktatoren eine Rolle spielen könnten: ein Versäumnis, das sogar vor dem Hintergrund der wenig beeindruckenden Standards der Technikfolgenabschätzung jener Zeit erstaunt.¹² Wenn es so weit ist, müssen wir darauf hoffen, dass wir nicht nur die technologische Kompetenz haben werden, eine Intelligenzexplosion einzuleiten, sondern auch die darüber hinaus erforderlichen Fähigkeiten besitzen, um die Detonation zu überleben.

Ehe wir uns jedoch dem zuwenden, was vor uns liegt, wird es sinnvoll sein, einen kurzen Blick auf die Geschichte der maschinellen Intelligenz von ihren Anfängen bis heute zu werfen.

Zeiten der Hoffnung und der Hoffnungslosigkeit

Im Sommer 1956 kamen am Dartmouth College zehn Wissenschaftler zu einem sechswöchigen Workshop zusammen, die sich alle für neuronale Netze, Automatentheorie und das Studium der Intelligenz interessierten. Dieses Treffen wird häufig als die Geburtsstunde des For-

schungsfelds der künstlichen Intelligenz betrachtet, und viele der Teilnehmer gelten heute als dessen Gründerväter. Ihr Optimismus kommt in einem Antrag an die Rockefeller-Stiftung zum Ausdruck, die Mittel für den Workshop bereitstellte:

Wir schlagen vor, dass eine zweimonatige, zehnköpfige Untersuchung der künstlichen Intelligenz durchgeführt wird. [...] Die Studie soll auf der Grundlage der Vermutung durchgeführt werden, dass jedes Merkmal des Lernens oder der Intelligenz überhaupt im Prinzip so genau beschreibbar ist, dass eine Maschine es simulieren kann. Es wird der Versuch gemacht werden, herauszufinden, wie man Maschinen baut, die Sprache verwenden, abstrahieren und Begriffe bilden, die Arten von Problemen lösen, welche derzeit den Menschen vorbehalten sind, und die sich selbst vervollkommen. 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 daran arbeitet.

Seit diesen kühnen Worten sind sechs Jahrzehnte vergangen, in denen sich auf dem Feld der künstlichen Intelligenz Begeisterungstürme und hohe Erwartungen mit Rückschlägen und Enttäuschungen abwechselten.

Die Zeit der anfänglichen Begeisterung, die mit dem Treffen in Dartmouth begann, wurde später von John McCarthy (dem Hauptorganisator der Veranstaltung) als die »Kein Trick!«-Ära bezeichnet. In jenen Tagen bauten die Forscher einfache Systeme, um die damals oft zu hörenden Behauptungen der Form »Keine Maschine wird jemals X können!« zu widerlegen. Diese Systeme konnten eine abgespeckte Version von X in einer »Mikrowelt« (einem wohldefinierten, begrenzten Bereich) ausführen, sodass die prinzipielle Machbarkeit einer Maschine, die X in der echten Welt tun könnte, nachgewiesen war. Ein solches frühes System, *Logic Theorist*, war in der Lage, die meisten Theoreme des zweiten Kapitels von Whiteheads und Russells *Principia Mathematica* zu beweisen, und kam sogar auf einen Beweis, der viel eleganter war als der ursprüngliche. Eine Weiterentwicklung des Programms, der *General Problem Solver*, war im Prinzip fähig, eine Vielzahl formal spezifizierter Probleme zu lösen.¹³ Dadurch war die Vorstellung, Maschinen könnten »nur numerisch denken«, widerlegt und gezeigt, dass Maschinen zur Deduktion und zum Erfinden logischer Beweise fähig sind.¹⁴ Außerdem schrieb man Programme zum Lösen von Analysis-Aufgaben für Erstsemester, von visuellen Analogieproblemen von der Art, wie sie