sollten wir als Spezies uns dennoch sehr gedemütigt fühlen [...] Diese neue Gefahr [...] ist sicherlich etwas, das uns ängstigen könnte.«

Als die Anti-Maschinisten in Erewhon »ehrliche Besorgnis für die Zukunft hegen«, sehen sie es als ihre »Pflicht, dem Übel zu steuern, so lange es noch in [ihrer] Macht steht«, und zerstören alle Maschinen. Turings Antwort auf die neue »Gefahr« und die »Angst« davor besteht darin, bei Bedarf »den Strom abzuschalten« (allerdings werden wir schon bald sehen, dass das nicht wirklich eine Option ist). In Frank Herberts Science-Fiction-Klassiker *Der Wüstenplanet* hat die Menschheit einer sehr fernen Zukunft knapp Butlers Djihad überlebt, einen katastrophalen Krieg gegen die »denkenden Maschinen«. Daraus entsteht ein neues Gebot: »Du sollst keine Maschine nach deinem geistigen Ebenbild machen.« Das schließt Rechenapparate jeder Art ein.

All diese drastischen Reaktionen spiegeln die innere Furcht vor Maschinenintelligenz wider. Ja, die Vorstellung superintelligenter Maschinen erregt Unbehagen. Ja, es ist rein logisch möglich, dass solche Maschinen die Welt übernehmen und sich die menschliche Rasse unterwerfen oder sie sogar auslöschen. Wenn wir allein danach gehen, ist die einzige vernünftige Reaktion, sofort jede Forschung an der künstlichen Intelligenz einzuschränken und insbesondere die Entwicklung und Verbreitung einer dem Menschen ebenbürtigen allgemeinen KI in welcher Form auch immer zu verbieten.

Wie die meisten KI-Forscher schaudert es mich bei diesem Gedanken. Wie kann es jemand wagen, mir vorzuschreiben, worüber ich nachdenken darf und worüber nicht? Jeder, der die Einstellung der KI-Forschung fordert, muss wirklich gute Argumente auf seiner Seite haben und viel Überzeugungsarbeit leisten. Das Ende der KI-Forschung würde nicht nur einen wichtigen Weg zu einem besseren Verständnis der menschlichen Intelligenz versperren, sondern auch eine großartige Möglichkeit verhindern, das Leben der Menschen zu verbessern und eine sehr viel bessere Zivilisation zu schaffen. Der wirtschaftliche Wert einer dem Menschen ebenbürtigen KI lässt sich mit Tausenden von Billionen US-Dollar beziffern. Konzerne und Regierungen dürften daher am Fortbestand der KI-Forschung sehr interessiert sein. Ihr Gewicht ist weit größer als das eines Philosophen, egal welchen »Ruf für besondere Gelehrsamkeit« dieser auch haben mag.

Ein weiterer Nachteil eines solchen Forschungsverbots ist die Schwierigkeit, es durchzusetzen. Fortschritte auf dem Weg zu einer allgemeinen KI werden meist auf den Tafeln und Whiteboards in Forschungseinrichtungen weltweit gemacht, und zwar bei der Aufstellung und Lösung mathematischer Probleme. Niemand weiß im Voraus, welche Ideen und Gleichungen man verbieten muss. Selbst wenn wir es wüssten, ist es kaum vernünftig, anzunehmen, dass ein solches Verbot durchsetzbar oder wirksam wäre.

Noch komplizierter wird die Sache dadurch, dass Forscher oft an anderen Dingen arbeiten und dabei indirekt zum Fortschreiten einer allgemeinen KI beitragen. Wie ich bereits gezeigt habe, führt die Forschung an spezifischen KI-Tools für Anwendungen wie Spiele, medizinische Diagnosen und Reiseplanungen oft zu Fortschritten bei Techniken, die sich auch bei vielen anderen Problemen einsetzen lassen und uns einer dem Menschen ebenbürtigen KI näherbringen.

Aus all diesen Gründen ist es eher unwahrscheinlich, dass die KI-Community oder die Regierungen und Konzerne, die über Gesetze und Forschungsmittel bestimmen, das Gorilla-Problem durch ein Verbot bestimmter KI-Forschungen lösen. Fakt ist: Wenn die einzige Lösung des Problems in diesem Weg besteht, wird es nicht gelöst werden.

Der Ansatz, der vermutlich am besten funktioniert, besteht darin, zu ergründen, warum die Erschaffung einer besseren KI vielleicht schlecht ist. Und die Antwort auf diese Frage kennen wir schon seit mehreren Tausend Jahren.

Das König-Midas-Problem

Norbert Wiener (den Sie aus Kapitel 1 kennen) hat in vielen Gebieten tiefe Spuren hinterlassen, darunter in der künstlichen Intelligenz, in den Kognitionswissenschaften und in der Kontrolltheorie. Anders als die meisten seiner Zeitgenossen interessierte er sich besonders für die Unvorhersehbarkeit komplexer Systeme in der Realität. (Er schrieb seine erste Abhandlung zu diesem Thema im Alter von zehn Jahren.) Er war überzeugt davon, dass das übertriebene Vertrauen der Wissenschaftler und Techniker in ihre eigenen Fähigkeiten, ihre Schöpfungen zu kontrollieren – gleich ob militärischer oder ziviler Natur –, katastrophale Folgen haben könnte.

bei der religiösen Zeitschrift *Primitive Expounder*, 1847 gegen mechanische Rechenmaschinen.[2]

> »Der Geist [...] läuft sich selbst davon und tut die Notwendigkeit der eigenen Existenz mit der Erfindung von Maschinen ab, die selbst *denken* [...] Doch wer weiß, ob solche Maschinen, wenn sie denn größere Vollkommenheit erreichen, nicht einen Plan ersinnen, all ihre eigenen Mängel zu beheben, und Ideen haben, die sich der sterbliche Geist nicht vorzustellen vermag!«

Dieser Text ist vielleicht die erste Überlegung hinsichtlich des existenziellen Risikos, das Rechenmaschinen darstellen – aber er ist recht unbekannt. Im Gegensatz dazu war Samuel Butlers 1872 veröffentlichtem Roman *Erewhon oder jenseits der Berge*, der das Thema sehr viel detaillierter aufgriff, sofortiger Erfolg beschieden. Erewhon ist ein Land, in dem alle mechanischen Apparate nach einem schrecklichen Bürgerkrieg zwischen Maschinisten und Anti-Maschinisten verbannt wurden. Ein Teil des Buchs mit dem Titel »Das Buch der Maschinen« erläutert die Ursprünge des Kriegs und stellt die Argumente beider Parteien vor.[3] Er stellt die Anfang des 21. Jahrhunderts neu aufgeflammte Debatte gespenstisch genau dar. Die Anti-Maschinisten führen vor allem die Behauptung ins Feld, dass Maschinen irgendwann so fortgeschritten sind, dass der Mensch die Kontrolle über sie verliert:

> »Schaffen wir uns nicht selbst in ihnen unsere Nachfolger als Herren der Erde, indem wir täglich die Schönheit und Feinheit ihrer Organisation vergrößern, täglich ihnen größere Geschicklichkeit verleihen und immer mehr von jener sich selbst regulierenden, selbstständigen Kraft mitteilen, welche mehr wert ist als jede Art von Geist? [...] Allein wir stehen vor der Alternative: entweder jetzt schwere Leiden erdulden zu müssen oder uns von unseren eigenen Geschöpfen allmählich überholt zu sehen, bis wir am Ende im Vergleich zu ihnen nicht höher rangieren als die Tiere des Feldes im Verhältnis zu uns.«[4]

Der Erzähler nennt auch das wesentliche Gegenargument der Pro-Maschinisten, das die Mensch-Maschine-Symbiose vorwegnimmt, auf die wir im nächsten Kapitel noch eingehen:

> »Nur von einer Seite wurde ein ernstlicher Angriff gegen sie unternommen. Der Autor dieser Gegenschrift argumentierte etwa so: Die Maschinen sind anzusehen als ein Teil der physischen Natur des Menschen; sie sind im Grunde nichts weiter als Glieder außerhalb des Körpers.«[5]

Zwar gewinnen die Anti-Maschinisten in Erewhon die Diskussion, aber Butler selbst scheint hin- und hergerissen zu sein. So beklagt er sich einerseits, dass Erewhonier seien »schnell bereit, den gesunden Menschenverstand auf dem Altar der Logik zu opfern, wenn ein Philosoph unter ihnen aufsteht, der sie durch seinen Ruf für besondere Gelehrsamkeit mitreißt« und sagt: »Sie sind bereit, sich für die Maschinen selbst zu opfern.« Andererseits beschreibt er eine erstaunlich harmonische, produktive und geradezu idyllische erewhonische Gesellschaft. Die Erewhonier erkennen die Torheit, sich erneut auf den Weg der mechanischen Intervention zu begeben, und stellen die Überreste der Maschinen in Museen aus »in der Weise eines Altertumsforschers, der sich für Druidensteine oder Pfeilspitzen aus Feuerstein interessiert«.

Alan Turing kannte den Roman. In einem Vortrag, den er 1951 in Manchester hielt, machte er sich Gedanken über die langfristige Zukunft der KI:[6]

> »Es erscheint wahrscheinlich, dass Maschinen, nachdem sie auf ihre Weise zu denken begonnen haben, nicht lange benötigen, um unsere geringe Leistung zu übertreffen. Maschinen sterben natürlich nicht. Und sie könnten sich miteinander unterhalten, um ihren Verstand zu schärfen. Irgendwann müssen wir damit rechnen, dass die Maschinen die Kontrolle übernehmen, wie Samuel Butler es in *Erewhon* zeigt.«

Im selben Jahr wiederholte Turing seine Bedenken in einer auf BBC 3 ausgestrahlten Radiosendung:

> »Wenn eine Maschine denken kann, dann könnte sie auch intelligenter denken, als wir es tun; und wo stehen wir dann? Selbst wenn wir es schaffen, die Maschinen in einer untergeordneten Stellung zu halten, indem wir sie beispielsweise im passenden Moment ausschalten, dann

ihre Vorgänger, denn eine kleine Armee von Robotern bringt ihr die Lagerboxen, aus denen sie die Ware nimmt. Doch gleichzeitig ist sie ein kleines Rad in einem großen Räderwerk, das von intelligenten Algorithmen gesteuert wird, die ihr vorgeben, wo sie zu stehen und welche Waren sie zu nehmen und abzufertigen hat. Sie steht nicht auf der Spitze der Pyramide, sondern ist schon halb darin begraben. Schon bald wird der Sand die Hohlräume der Pyramide gefüllt und ihre Rolle überflüssig gemacht haben.

5

ÜBERMÄßIG INTELLIGENTE KI

Das Gorilla-Problem

Es braucht nicht viel Vorstellungskraft, um zu erkennen, dass es vielleicht keine so gute Idee ist, etwas zu erschaffen, das klüger ist als man selbst. Wir wissen, dass unsere »Herrschaft« über unsere Umwelt und andere Spezies ein Resultat unserer Intelligenz ist. Allein der Gedanke, dass etwas – ob Roboter oder Außerirdischer – existieren könnte, das intelligenter ist als wir, verursacht bei uns ein mulmiges Gefühl. Vor etwa zehn Millionen Jahren erschufen die Vorfahren der heutigen Gorillas (zugegebenermaßen unabsichtlich) die genetische Linie, aus der der moderne Mensch entstammt. Was halten die Gorillas davon? Wenn sie uns ihre Meinung zu ihrer aktuellen Situation im Vergleich zu der des Menschen mitteilen könnten, dürfte diese ohne Frage sehr negativ ausfallen. Ihre Gattung hat praktisch nur die Zukunft, die wir ihr zugestehen. Wir möchten uns den superintelligenten Maschinen gegenüber bestimmt nicht in der gleichen Lage befinden. Ich nenne dies das *Gorilla-Problem*. Es geht darum, ob die Menschen ihre Überlegenheit und Autonomie in einer Welt beibehalten können, in der es auch Maschinen mit einer erheblich höheren Intelligenz gibt. Charles Babbage und Ada Lovelace, die 1842 die Analytical Engine konstruierten und programmierten, waren sich des Potenzials dieser Maschine bewusst, hatten aber, soweit wir wissen, diesbezüglich keinerlei Bedenken.[1] Allerdings wetterte Richard Thornton, Herausge-