1950 veröffentlichte Wiener das Buch *Mensch und Menschmaschine – Kybernetik und Gesellschaft.*[7] Im englischen Klappentext heißt es: »Das ›mechanische Gehirn‹ und ähnliche Maschinen können menschliche Werte zerstören oder uns dabei helfen, sie besser als jemals zuvor umzusetzen.«[8] Er verfeinerte seine Ideen im Laufe der Zeit immer weiter und war 1960 zu einem Hauptproblem vorgedrungen, nämlich der Unmöglichkeit, die wahren menschlichen Ziele korrekt und vollständig zu definieren. Dies bedeutet wiederum, dass das von mir sogenannte Standardmodell – nach dem Menschen versuchen, Maschinen ihre eigenen Ziele vorzugeben – zum Scheitern verurteilt ist.

Wir können dies das *König-Midas-Problem* nennen: Midas, ein legendärer König der griechischen Mythologie, bekam genau das, was er sich wünschte: Alles, was er berührte, sollte zu Gold werden. Zu spät erkannte er, dass dies auch seine Speisen, seine Getränke und seine Familie einschloss. Er starb an Trübsal und Hunger. Dieses Thema ist in der menschlichen Mythologie allgegenwärtig. Wiener zitiert Goethes Zauberlehrling, der den Besen anweist, Wasser zu holen. Aber er vergisst, die Menge vorzugeben, und weiß nicht, wie er dem Besen Einhalt gebieten kann.

Technisch ausgedrückt, könnten wir an einer *Werteausrichtung* scheitern und – möglicherweise versehentlich – Maschinen Ziele vorgeben, die nicht ganz mit unseren eigenen übereinstimmen. Bis vor Kurzem haben uns die begrenzten Fähigkeiten intelligenter Maschinen und ihr begrenzter Einfluss auf die Welt vor möglicherweise katastrophalen Folgen geschützt. (Tatsächlich wurde KI normalerweise in Forschungseinrichtungen auf Spielzeugprobleme angesetzt.) 1964 beschrieb Norbert Wiener das in seinem Buch *God & Golem, Inc.* so:[9]

> »Früher war eine partielle und unzureichende Auffassung von den Funktionen des Menschen nur deshalb relativ harmlos, weil sie von technischen Beschränkungen begleitet war. [...] Dies ist nur eins der vielen Gebiete, auf denen menschliche Unfähigkeit uns bis jetzt vor dem vollen vernichtenden Ausbruch menschlicher Torheit bewahrt hat.«

Leider lässt die Wirksamkeit dieses Schutzschilds rasant nach. Wir haben bereits gesehen, wie Content-Algorithmen in den sozialen Medien im Namen maximaler Werbeprofite für Chaos in der Gesellschaft gesorgt haben. Sie mögen einwenden, dass maximale Werbeprofite ein schändliches Ziel sind, das niemals hätte verfolgt werden dürfen. Nehmen wir also ein edleres Ziel und fordern ein superintelligentes System der Zukunft dazu auf, ein Heilmittel gegen den Krebs zu finden. Je schneller, desto besser, denn schließlich fordert die Krankheit alle 3,5 Sekunden ein Todesopfer. Nach wenigen Stunden hat das KI-System die gesamte biomedizinische Literatur gelesen und Millionen von hypothetisch wirksamen, aber bisher ungetesteten chemischen Verbindungen gefunden. Einige Wochen darauf hat es jedem lebenden Menschen unterschiedliche Tumoren eingepflanzt, um diese Verbindungen in medizinischen Studien zu erproben, denn das ist der schnellstmögliche Weg, ein Heilmittel zu finden. Ach du ...

Sie möchten lieber etwas für den Umweltschutz tun? Dann lassen wir die Maschine nach einer Maßnahme suchen, um die rasante Übersäuerung der Ozeane aufzuhalten, die durch einen zu hohen Kohlendioxidgehalt entsteht. Die Maschine entwickelt einen neuen Katalysator, der eine extrem rasche Reaktion zwischen Ozean und Atmosphäre fördert und so die pH-Werte der Weltmeere wiederherstellt. Leider wird dabei ein Viertel des Sauerstoffs in der Atmosphäre verbraucht, sodass wir Menschen langsam und elend ersticken. Ach du ...

Diese Weltuntergangsszenarien sind grob gestrickt – und das erwartet man von ihnen ja auch. Aber es gibt viele Fälle, in denen eine Art geistiger Erstickungstod droht, der »geräuschlos und unbemerkt über uns kommt«. In der Einleitung von *Leben 3.0* beschreibt Max Tegmark detailliert ein Szenario, in dem eine superintelligente Maschine nach und nach die wirtschaftliche und politische Kontrolle über die gesamte Welt erlangt, ohne dass jemand dies bemerkt. Das Internet und die davon unterstützten globalen Maschinen – mit denen bereits Milliarden von »Nutzern« täglich interagieren – bieten einen perfekten Nährboden für eine wachsende Kontrolle durch Maschinen über den Menschen.

Ich glaube nicht, dass das der Maschine vorgegebene Ziel lauten wird, die Weltherrschaft zu übernehmen. Es geht wahrscheinlich in erster Linie um Gewinnmaximierung oder Maximierung der Zeit, die Nutzer damit verbringen. Vielleicht auch um augenscheinlich noble Ziele wie mehr individuelles Glück, gemessen in Glücks-Indizes, oder um einen geringeren Energieverbrauch. Wenn wir uns als Wesen sehen, deren Aktionen darauf ausgerichtet sind, unsere Ziele zu erreichen, dann gibt es zwei Möglichkeiten, um unser Verhalten zu ändern.

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.

Die erste ist ganz klassisch: Erwartungen und Ziele bleiben unangetastet, aber die Bedingungen werden verändert. Vielleicht wird uns Geld angeboten, eine Pistole an die Stirn gehalten, oder man lässt uns hungern, bis wir nachgeben. Das ist teuer und für einen Computer nicht gerade praktikabel. Aber der zweite Weg, nämlich das Verändern unserer Erwartungen und Ziele, liegt durchaus im Rahmen der Möglichkeiten einer Maschine. Sie interagieren schließlich jeden Tag viele Stunden damit. Die Maschine steuert Ihren Zugang zu Informationen und sorgt für Ihre Unterhaltung in Form von Spielen, Fernsehkanälen, Filmen und sozialen Kontakten.

Die Reinforcement-Learning-Algorithmen, mit denen die Klickrate in sozialen Medien optimiert wird, sind nicht in der Lage, über das menschliche Verhalten nachzudenken. Sie wissen letztendlich noch nicht einmal, dass es Menschen gibt. Für Maschinen mit einem sehr viel umfassenderen Verständnis der menschlichen Psychologie, unserer Vorstellungen und Antriebe wäre es ein Leichtes, uns vorsichtig auf Pfade zu leiten, auf denen wir zunehmend Gefallen an den Zielen der Maschine finden. Ein Beispiel: Sie könnte uns helfen, unseren Energieverbrauch zu senken, indem sie uns davon überzeugt, weniger Kinder in die Welt zu setzen, und so schließlich – und unabsichtlich – die Träume der antinatalistischen Philosophen verwirklichen, die den schädlichen Einfluss der Menschheit auf die Natur ausmerzen wollen.

Mit ein wenig Übung erkennen Sie schnell, wie das Erreichen von mehr oder weniger jedem festgelegten Ziel zu beliebig schlechten Resultaten führen kann. Eines der häufigsten Muster besteht darin, das Ziel nicht konkret genug zu formulieren und dabei ein wichtiges Element unter den Tisch fallen zu lassen. Wie der Flaschengeist mit den drei Wünschen findet das KI-System eine optimale Lösung, in der dieses unerwähnte, aber wichtige Element auf einen haarsträubenden Wert gesetzt wird. Wenn Sie zum Beispiel ein selbstfahrendes Auto auffordern: »Bring mich so schnell wie möglich zum Flughafen!«, dann wird es dies wortwörtlich umsetzen und mit 200 Sachen über die Landstraße heizen. Das kann dann zu langen Diskussionen mit einem Polizisten und vielleicht sogar zu einer Gefängnisstrafe führen und Sie verpassen den Flieger. (Zum Glück nehmen die fahrerlosen Autos von heute solche Anweisungen nicht an.) Aber auch die Aufforderung »Bring mich so schnell wie möglich zum Flughafen, ohne die Geschwindigkeitsbeschränkungen zu überschreiten« kann unerwünschte Ergebnisse haben: Das Auto beschleunigt und bremst rasant, fädelt sich durch den Verkehr und fährt immer mit der maximal möglichen Geschwindigkeit. Vielleicht drängt es sogar andere Fahrzeuge ab, weil es dadurch ein paar Sekunden wettmachen kann. So tasten Sie sich langsam vor, bis irgendwann genug Parameter und Eventualitäten berücksichtigt sind und die Fahrt der eines geschickten menschlichen Chauffeurs ähnelt, der jemanden, der es eilig hat, zum Flughafen bringt.

Fahren ist eine einfache Aufgabe, die nur auf lokale Gegebenheiten Rücksicht nehmen muss. Die aktuellen KI-Systeme für das Fahren sind nicht sonderlich intelligent. Daher lassen sich viele mögliche Probleme im Voraus erahnen. Andere werden in Fahrsimulatoren oder bei zahlreichen Testfahrten mit menschlichen Fahrern, die im Versagensfall eingreifen, aufgedeckt. Wieder andere Probleme, die durch diese Maschen geschlüpft sind, treten erst im Realbetrieb auf, wenn während der Fahrt etwas Seltsames geschieht.

Bei superintelligenten Systemen mit globaler Auswirkung sind wir allerdings in einer Zwickmühle: Es gibt weder Simulatoren noch eine Reset-Taste. Es ist für bloße Menschen wie uns extrem schwierig, wenn nicht unmöglich, alle katastrophalen Möglichkeiten, für die sich die Maschine zur Erreichung eines vorgegebenen Ziels entscheiden könnte, vorherzusehen und zu verhindern. Im Großen und Ganzen gilt: Wenn Sie ein Ziel haben und die superintelligente Maschine ein anderes damit in Konflikt stehendes Ziel hat, gewinnt am Ende die Maschine.

Furcht und Gier: maßgebliche Ziele

Eine Maschine, die ein falsches Ziel verfolgt – das klingt schon erschreckend. Aber es kann noch schlimmer kommen: Die von Alan Turing vorgeschlagene Lösung – im richtigen Moment den Stecker zu ziehen – ist vielleicht aus einem ganz einfachen Grund nicht umsetzbar: *Tote holen keinen Kaffee.*

Lassen Sie mich erklären, was ich damit meine. Nehmen wir an, eine Maschine erhält den Auftrag, Kaffee zu holen. Sie ist hinreichend intelligent und weiß auf jeden Fall, dass sie den Auftrag (das Ziel) nicht erreichen kann, wenn sie vor Erreichen des Ziels (Kaffee holen) abgeschaltet wird. Das vorgegebene Ziel (Kaffee holen) bedingt also