Katharina
Zweig

Weiß die KI, dass sie nichts weiß?

Katharina
Zweig

Weiß die KI, dass sie nichts weiß?

Wofür wir Chatbots und KI-Agenten nutzen sollten, wo sie sich irren und wo wir aufpassen müssen

ChatGPT und Co. einfach erklärt

HEYNE

Penguin Random House Verlagsgruppe FSC® N001967

2. Auflage
Originalausgabe 09/2025

produktsicherheit@penguinrandomhouse.de
(Vorstehende Angaben sind zugleich
Pflichtinformationen nach GPSR.)

www.heyne.de

Redaktion: Evelyn Boos-Körner
Abbildungen: Abbildung 8,14 und 15: Franz Hoegl,
alle anderen Abbildungen: Katharina Zweig
Umschlaggestaltung: Favoritbüro
Satz: satz-bau Leingärtner, Nabburg
Druck und Bindung: GGP Media GmbH, Pößneck
Printed in Germany
ISBN: 978-3-453-21907-6

Inhalt

KAPITEL 1

Wenn Worte meine Sprache wären …

»Mir fehlen die Worte, ich hab die Worte nicht …
Ich bin ohne Worte, ich finde die Worte nicht …
Ich hab keine Worte für dich …
Ich hab die Worte nicht, dir zu sagen, was ich fühl«

Tim Bendzko, »Wenn Worte meine Sprache wären«,
vom Album *Wenn Worte meine Sprache wären*, 2011

Uns wird viel versprochen, wenn es um die großen Sprachmodelle geht. Haben Sie nicht auch schon davon gehört, dass ChatGPT Texte zusammenfassen oder vergleichen kann, dass es unsere Anweisungen verstehen, beurteilen und bewerten kann? Das neue große Ding ist, dass Menschen Sprachmodelle als KI-Agenten nutzen sollen, also Software, die in unserem Namen handeln soll: Man sagt der Maschine beispielsweise, dass man eine Reise nach Island machen will, nennt die Daten und dann soll der KI-Agent **selbstständig** alle dafür notwendigen Buchungen durchführen. Dafür fragt sich die Maschine zuerst selbst: »Was ist alles notwendig, um eine Reise nach Island zu machen?« Im Anschluss arbeitet sie dann die einzelnen Schritte der Reihe nach ab. Würden Sie einer solchen KI Ihre Kreditkartendaten für die Reisevorbereitung übergeben? Ich rate dringend davon ab! Denn dafür müsste die Maschine Sie verstehen können und über Ihren Auftrag nachdenken können – nur dann können Sie ihr vertrauen, dass sie die richtigen Schritte ausführen wird.

Aber worüber sprechen wir hier eigentlich, wenn wir den großen

Sprachmodellen zutrauen, dass sie etwas verstehen, beurteilen, bewerten, vergleichen und zusammenfassen oder gar an unserer Stelle handeln können?

Ich glaube, dass uns als Menschen hier die Worte fehlen, um zu beschreiben, was ChatGPT und andere große Sprachmodelle eigentlich tun. Und auch wenn Tim Bendzko mit seinem Lied »Wenn Worte meine Sprache wären«[1] die Wortlosigkeit angesichts einer großen Liebe besang, passen die Ausschnitte aus seinen Lyrics für mich auch angesichts des großen Umbruchs, den diese neue Technologie mitbringt – und wie sich das für uns anfühlt. Denn wie sollen wir es nennen, was diese Maschinen tun können?

Eines meiner ersten größeren Experimente mit ChatGPT fand in den Winterferien 2022 statt. Zu diesem Zeitpunkt hatte ich damit schon ein bisschen herumgespielt, Informationen zu meinem eigenen Namen abgefragt und auch schon einige Interviews zu den möglichen Auswirkungen geführt.

In den Weihnachtsferien wollte ich mich näher damit beschäftigen und als meinen *partner in crime* habe ich natürlich meine Tochter ausgewählt. »Schatz, das sollten wir uns mal gemeinsam angucken. Angeblich kann man damit auch Hausaufgaben erledigen!« Meine Tochter hatte aber, entgegen meiner Erwartung, so gar keine Lust dazu. »Nee, Mama, ich will meine Hausaufgaben selbst machen.« Aha? Das hat mich auf der einen Seite sehr gefreut, pädagogisch wertvoll und so. Aber wenigstens einmal gucken, was man mit der neuen Technologie machen könnte, sollte doch wohl drin sein, oder? »Und wenn wir damit mal den Dankesbrief für dein Weihnachtsgeschenk von Tante Franja schreiben lassen?«, lockte ich sie, wohl wissend, dass das Schreiben von Briefen für sie keine geliebte Tätigkeit war. »Na gut.«

Mein Auftrag an die Maschine ergab einen durchaus lesenswerten Brief, sehr höflich und mit allen Informationen zum Geschenk, die wir dem Computer mitgegeben hatten. Interessanterweise wurde Franja aber von der Maschine gesiezt. Das ging natürlich nicht, schließlich würde meine Tochter ihre Patentante niemals siezen. Daher bat ich die Software, den Brief noch einmal so umzuschreiben, dass Franja

überall geduzt würde. Zu meinem großen Erstaunen konnte die Maschine das! Das ist ja nun keine kleine Anpassung: Erst müssen alle »Sie«, »Ihnen«, »Ihre« gefunden werden und durch entsprechende »du«, »dir«, »deine« ersetzt werden. Aber ohne eine entsprechende Anpassung der Verben wäre das Ergebnis immer noch falsch. Ich war daher sehr skeptisch, was diese komplexe Aufgabe anging. Aber die Maschine hat die erbetene Änderung ganz anstandslos und ohne Fehler umgesetzt: Vor mir lag ein sehr schöner Entwurf für einen Dankesbrief an unsere Freundin. Hatte die Maschine mich **verstanden**? Das Ergebnis ließ eigentlich gar keinen anderen Schluss zu – schließlich würde ich von einem Menschen, der diesen Auftrag ohne Fehler durchführt, doch auch sagen, dass er mich verstanden hat, oder? Und ich würde diesem Menschen eine gehörige Portion Deutschkenntnisse zusprechen und damit auch ein Mindestmaß an Intelligenz.

Ist eine solche Maschine, die meinen Auftrag korrekt ausführt, intelligent, kann sie denken?[2]

Darum geht es in diesem Buch: Erstens um die Frage, was Sprachmodelle wie ChatGPT, Claude, LaMDA, Gemini, Perplexity AI oder Llama können und wie wir sprachlich genauer erfassen, was sie tun – und was nicht.

Zweitens möchte ich Ihnen helfen, eine Intuition dafür zu entwickeln, was Sprachmodelle zuverlässig erledigen können. Drittens werde ich argumentieren, warum man basierend auf dieser Technologie keine KI-Agenten losschicken sollte, um echte Dinge in der echten Welt für Menschen zu erledigen.

Bezüglich dieser Frage gibt es unter den Expertinnen und Experten eine kleinere Fraktion, die alles für das Ergebnis reiner Statistik hält – sie bezeichnen Maschinen wie ChatGPT als *stochastic parrots*, also »stochastische Papageien«, die einfach vor sich hin plappern; im Deutschen würden wir eher vom »Nachäffen« sprechen. Diese Gruppe vertritt die Meinung, dass auch die großen Sprachmodelle

nur wie plappernde Papageien gelernt hätten, wann Menschen welche Wörter in welchem Kontext sagen, und dies wiederholen können. Die deutlich größere Fraktion von Experten sieht in den Sprachmodellen die ersten Anflüge von Nachdenken, wie hier z. B. im Juni 2023 Sam Altman, der als Geschäftsführer von OpenAI natürlich auch seine eigenen Ziele verfolgt: »Ist (die Idee der *stochastic parrots*) immer noch eine weit verbreitete Ansicht? Ich meine, wird das so gesehen – gibt es immer noch viele vernünftige Personen, die so denken? Mein Eindruck ist, dass die Leute nach GPT-4 größtenteils aufgehört haben, das zu sagen, und stattdessen angefangen haben zu sagen: ›Okay, es funktioniert, aber es ist zu gefährlich.‹«

Abbildung 1: Streit in der Wissenschaft um Sprachmodelle: Die einen halten sie für rein statistisch arbeitende Software, die anderen sehen erste Anzeichen von echtem Nachdenken.

Altman wird auch damit zitiert, dass das Sprachmodell GPT-4, die damals neueste Variante, »in gewissem Maße« nachdenken könne[3] – er nutzt dabei das Wort *reasoning*, das im Englischen eine große Palette von Aspekten abdeckt: das logische Schließen in der Mathematik, das Nachdenken über einen Sachverhalt, Begründen, Analysieren,

die begründete Schlussfolgerung. Altman vertritt die Ansicht, dass die sogenannte Künstliche Intelligenz schon in ein paar Jahren zur Superintelligenz geworden sein könnte, mit der sowohl das Klimaproblem »gefixt« als auch eine Marskolonie aufgebaut werden könne und nicht zuletzt alle Regeln der Physik entdeckt werden könnten.[4] Er gehört damit zu der Gruppe von Experten, die glauben, dass wir kurz vor der Entwicklung von Maschinen mit übermenschlichen intellektuellen Fähigkeiten stehen. Man könnte die eine Gruppe die »Nachäffer«-Fraktion und die andere die »Intelligenzbestien«-Fraktion nennen. Die beiden Wörter sind nicht zufällig aus dem Bereich der Tierwelt gewählt: Tiere kommunizieren untereinander und mit uns – aber es ist nicht ganz einfach zu untersuchen, wie viel sie wirklich verstehen und welches Verhalten einfach nur gelernt ist. Die Menschheit braucht daher neue Methoden, um zu untersuchen, inwieweit Maschinen etwas »verstehen« oder »schlussfolgern« können.

Es ist sicherlich kein großer Spoiler, wenn ich schon jetzt verrate, dass die Maschinen keine der heute vielfach versprochenen Fähigkeiten im menschlichen Sinne vollständig beherrschen: Sie können weder zusammenfassen noch vergleichen, können uns weder verstehen noch beurteilen. Und doch können sie Teile davon: Oft erledigen sie unsere Aufträge so, als ob sie uns verstünden; schreiben Texte, die wie Zusammenfassungen oder Vergleiche aussehen; verbessern uns, wo wir selbst den Fehler nicht gesehen hätten; nennen Fakten, die richtig sind; schreiben eine Analyse, die sich sinnvoll anhört; oder unterteilen eine große Aufgabe sinnvoll in kleinere Aufgaben.

In diesem Buch arbeite ich heraus, worin die Unterschiede zwischen menschlichen Tätigkeiten und denen von Sprachmodellen bestehen, und diskutiere, wie wir über diese Tätigkeiten reden könnten, um die verbleibenden Unterschiede zu bezeichnen. Und dazu schlage ich vor, diese Tätigkeiten von Maschinen, die menschlichen Tätigkeiten ähneln, mit einer ~ (Tilde) zu kennzeichnen. Dieses Zeichen wird in den Naturwissenschaften häufig verwendet, um darzustellen, dass etwas annähernd oder bis auf einige Faktoren dasselbe ist. Dann könnte man sagen: Der Computer ~versteht mich, ~fasst Texte ~zusammen und

schreibt ~Bewertungen. Nichts davon ist leistungsgleich zu dem, was der Mensch tut, das werde ich im Folgenden erklären, aber das meiste ist auch nicht extrem weit weg davon: Es enthält Anteile von dem, was Menschen von anderen Menschen erwarten, wenn sie von verstehen, zusammenfassen, vergleichen, bewerten und beurteilen sprechen. Damit bin ich für mein Buch auf jeden Fall schon einmal ausgerüstet und habe Worte, um zu sagen, »was ich fühl«: um zu beschreiben, wie der Stand der Dinge ist. Sie werden daher im Folgenden immer wieder einmal sehen, dass ich zwischen der menschlichen Durchführung und der maschinellen Durchführung in dieser Form unterscheide. **Denn eines ist klar: Wir brauchen sprachliche Differenzierungen der Tätigkeiten, die von Mensch und Maschine nicht vollständig gleich durchgeführt werden. Denn nur wenn wir verstehen, was die Maschine kann, können wir sie optimal einsetzen** – und dafür habe ich dieses Buch geschrieben.

Daher nehme ich Sie mit auf die Reise, die ich in den letzten 24 Monaten seit Veröffentlichung von ChatGPT-3 unternommen habe, um mir selbst einen Überblick über die Technologie zu verschaffen. Ich werde dabei viele Gebiete streifen: Kommunikation zwischen Menschen, das vorhersagende Gehirn, die Entstehung von Weltmodellen, den Konstruktivismus, das chinesische Zimmer von Searle, den Bender-Koller-Oktopus und vieles mehr. Ich hoffe, dass Sie Lust auf diese intellektuelle Reise haben, mit der Sie einen großen Überblick bekommen über das, was gerade diskutiert wird. Damit es nicht zu anstrengend wird, habe ich die Reise für Sie in 33 Abschnitte unterteilt. Wenn Sie jeden Tag einen Abschnitt lesen, haben Sie in einem Monat einen guten Überblick darüber, was Sprachmodelle können und was nicht. Jeder inhaltliche Abschnitt bekommt auch eine Zusammenfassung – wenn's Ihnen einmal zu detailliert wird, »hüpfen« Sie am besten direkt dorthin – die Zusammenfassungen sind so geschrieben, dass man den Teil, der folgt, nachvollziehen kann.

Ich hoffe, dass Sie die Entdeckungsreise genauso spannend finden werden wie ich, als ich die einzelnen Wegpunkte für mich, für dieses

Buch und damit für Sie, meine lieben Leserinnen und Leser, recherchiert habe. Ich persönlich mag es, das *big picture* zu verstehen, die Art und Weise, wie Dinge zusammenhängen. Natürlich bin ich für ein paar der angrenzenden Gebiete der Sprachwissenschaften und Semantik (der Bedeutung von Symbolen, insbesondere auch von Wörtern) und der Kognition keine Expertin – hier habe ich mich bestmöglich eingelesen, aber sicherlich nicht immer den gesamten Stand der Wissenschaft erfasst. Ich habe diejenigen Modelle, Studien und Erkenntnisse ausgewählt, die sich für mich zu einem kohärenten Gesamtbild zusammenfügen. Ich baue darauf, dass diejenigen unter Ihnen, die jeweils in diesen Gebieten Experten und Expertinnen sind, mir den ein oder anderen groben Strich verzeihen und die feineren Details in ihren eigenen Texten hinzufügen. Ein solches Buch kann nicht fehlerfrei sein, aber da keine Expertin und kein Experte in allen beteiligten Feldern gleich erfahren sein kann, braucht es jemanden, der die groben Zusammenhänge einmal erfasst. Nur dann können wir als Gesellschaft verstehen, was diese Technologie verspricht und was sie einhalten wird.

Denn eines ist klar: Sprachmodelle wie ChatGPT, Llama, Perplexity AI und weitere darauf basierende Software werden nicht mehr verschwinden und sie werden, wenn sie bestmöglich eingesetzt werden, unglaublich viel verändern. Aber wenn sie schlecht eingesetzt werden, wird es lange dauern, die schlechten Effekte zu entdecken und zu vermeiden – daher lassen Sie uns gemeinsam auf die Reise gehen und damit unnötige und schädliche Einsätze vermeiden und den besten Einsatz ermöglichen.

Teil I führt in die KI ein, aber auch in die Grundlagen der Kommunikation zwischen Menschen, soweit diese für das Buch notwendig sind.

Teil II steigt dann in den Maschinenraum der Technologie – denn was eine Maschine kann, sieht man am besten, wenn man sich mit der Technologie beschäftigt.

Teil III zieht die Schlussfolgerungen aus den ersten beiden Teilen und vermittelt Ihnen eine Intuition und einen Prozess dafür, zu entscheiden, wann Sie sprachmodellbasierten Systemen, insbesondere den sogenannten KI-Agenten, vertrauen können. Ihre Kreditkarte gehört dabei nie zu den Dingen, die Sie diesen Systemen anvertrauen können. Sollten Sie bei Teil II irgendwo hängenbleiben, ist Teil III auch unabhängig davon verständlich. Viel Spaß beim Lesen!

TEIL I

Künstliche Intelligenz und Sprache

KAPITEL 2

Auftakt: Das Phänomen der textenden Kisten

Jeder, der schon einmal mit den neuen Chatbots herumgespielt hat, ist erst einmal erstaunt, was die Maschinen können: Es fühlt sich oft so an, als hätte man einen völlig vernünftigen Gesprächspartner auf der anderen Seite. Nach einer Weile beginnt man dann, schwierigere Fragen zu stellen und erwischt die Maschine bei der ein oder anderen **Halluzination**.

Zum Beispiel hat die Maschine den Journalisten Hilmar Schmundt als rechte Hand von Hitler bezeichnet, weil sie ihn mit seinem Onkel ~verwechselt hat.[5] Dem australischen Politiker Brian Hood ~dichtete die Maschine ~an, er sei wegen Korruption verurteilt worden und habe 30 Monate im Gefängnis verbracht – dabei war er der Whistleblower, der die Korruption zur Anklage gebracht hatte.[6] Der amerikanische Rechtsanwalt Steven A. Schwartz, der sich gegen ein Urteil eines Gerichts wehren wollte, suchte bei der Verfassung der Gegenschrift Hilfe von ChatGPT. ChatGPT ~erfand aber kurzerhand einfach die Fälle, auf denen die Argumentation basierte. Der Anwalt stutze zwar kurz, als er die genannten Fälle nicht finden konnte, aber schlussendlich hielt er die Maschine für schlauer als sich selbst – das kostete die Kanzlei am Ende 5000 US-Dollar Strafe und einiges an Reputation.[7]

Tatsächlich hat es sich eingebürgert, bei solchen Fehlern von Sprachmodellen von Halluzinationen zu sprechen, obwohl es sich dabei um einen ungeeigneten Begriff handelt. Die Psychologin Alessia McGowan und ihre Co-Autoren weisen darauf hin, dass es bei Halluzinationen um eine Wahrnehmungsstörung geht: Halluzinationen sind als real wahrgenommene Sinneseindrücke, für die es keinen erkenn-

baren äußeren Reiz gibt. Die Psychologen schlagen daher den Begriff der **Konfabulation**[8] vor, dem Erfinden von Inhalten, die ein Patient in diesem Moment für wahr hält. Hierbei handelt es sich um Störungen der Spracherzeugung, Sprache wird also nicht so erzeugt, wie es normalerweise der Fall ist. Daher bleibe ich im Folgenden bei dem Wort **Konfabulation**, also dem Generieren von Texten, die eindeutig falsch sind.

Abbildung 2: Beim Menschen bezeichnet man Wahrnehmungsstörungen als Halluzination, das Erfinden von Antworten ohne inhaltliche Substanz dagegen als Konfabulation. Bei Sprachmodellen sollte man daher von Konfabulation reden, auch wenn sich der Begriff Halluzination eingebürgert hat.

Diese Fälle beziehen sich auf die alte Version, ChatGPT-3.5 der Firma OpenAI, die relativ viele Fehler gemacht hat. Aber die neue Version namens ChatGPT-4o, mit o für *omni* (lateinisch für *alles*), ist wirklich erstaunlich. Ich kann Sie nur einladen, sich die Videos auf der Webseite von OpenAI anzusehen.[9] Dort können Sie miterleben, wie zwei KI-Systeme miteinander sprechen, und das funktioniert so: Ein menschlicher

Moderator sitzt an einem Tisch mit zwei Handys; auf jedem ist eines der beiden KI-Systeme installiert. Bei dem ersten Handy läuft die Kamera und das KI-System übersetzt, was es »sieht«, in Text und dann in Sprache. Das andere System auf dem zweiten Handy »hört« zu und stellt Fragen.

Dabei kommt zwar nicht viel mehr heraus, als dass der von der Videokamera gefilmte menschliche Moderator eine schicke schwarze Lederjacke anhat und sehr modern aussieht, aber dabei zuzuhören, ist trotzdem einfach unglaublich. Am Ende bittet der Moderator die beiden Systeme, aus dem gemeinsam »Erlebten« einen Song zu kreieren – wobei jedes System immer nur eine Zeile singen soll. Und dann singen sich die zwei KI-Systeme an, über den schönen modernen Raum und die schwarze Le-he-heder-ja-hacke! Das ist nicht schön und sicher nicht kunstvoll, aber absolut neuartig. Beeindruckend ist, dass beide Systeme fast so schnell antworten wie Menschen, obwohl eine Vielzahl von Rechenschritten durchlaufen werden müssen. Es verwundert dann kaum noch, dass das System laut Marketing von OpenAI auch bei den Mathehausaufgaben helfen oder als Simultanübersetzer zwischen Englisch und Spanisch dienen kann. Mich persönlich hat am

meisten überzeugt, wie ChatGPT zusammen mit der Software Be-MyEyes blinden Personen helfen kann: Das System nimmt die Umgebung mit der Kamera wahr und antwortet dann auf Fragen. Im Video wird gezeigt, wie ein Mann das System bittet, den entgegenkommenden Verkehr auf ein Taxi zu scannen, das keine Fahrgäste hat, damit er es heranwinken kann. Und das gelingt tatsächlich. Auch wenn ich nicht zu den Personen gehöre, die schnell zu Superlativen greifen: Das ist wirklich extrem beeindruckend.

Aber zeigt es auch, dass die Maschinen denken können? Sogar übermenschliche Fähigkeiten haben und bald die sogenannte allgemeine Künstliche Intelligenz ausbilden? Damit ist gemeint, dass eine Maschine alle Aufgaben, die der Mensch lösen kann, auch selbst lösen kann – es wird vermutet, dass sie dazu auch ein Bewusstsein entwickeln müsste. Es sind diese Systeme, vor denen unter anderem der Physiker Stephen Hawking, Nobelpreisträger Geoffrey Hinton, Elon Musk und Wissenschaftsphilosophen wie Nick Bostrom und Max Tegmark warnen. Letzterer veröffentlichte mit seinem Future of Life Institute, das versucht, die Risiken von Technologie zu reduzieren, einen öffentlichen Brief. Darin wird gewarnt vor der Weiterentwicklung von KI-Systemen, die leistungsstärker als GPT-4 sind.[10] GPT-4 bezeichnet das grundlegende Sprachmodell, auf dem die Chatbot-Variante ChatGPT-4 beruht. In der Petition wird behauptet, dass moderne KI-Systeme schon heute bei der Lösung allgemeiner Aufgaben mit dem Menschen konkurrieren könnten und man sich deshalb fragen müsste: »Sollten wir nicht-menschliche Intelligenzen entwickeln, die uns schließlich zahlenmäßig übertreffen, überlisten, überflüssig machen und ersetzen könnten? Sollten wir den Verlust der Kontrolle über unsere Zivilisation riskieren?« Es wurde auch eine freiwillige Pause oder gar eine regierungsseitig verordnete Zwangspause der weiteren Entwicklung (ein Moratorium) gefordert. Die Autoren vermerken: »Dies bedeutet keine generelle Pause der KI-Entwicklung, sondern lediglich eine Abkehr vom gefährlichen Wettlauf zu immer größeren, unberechenbaren Blackbox-Modellen mit emergenten Fähigkeiten.« Die ersten fünf Unterzeichner sind die bekannten KI-Forscher Yoshua

Bengio und Stuart Russell, Elon Musk[11], der Apple-Co-Gründer Steve Wozniak und der Historiker Yuval Noah Harari.

Die ersten Meldungen, dass die neuen Sprachmodelle »viel zu gefährlich« seien, um sie überhaupt öffentlich zu machen, kamen aber schon weit vor der vierten Version des Sprachmodells GPT auf. Ehrlich gesagt wurden KI-Systeme schon seit Jahrzehnten für gefährlich gehalten – und das hat mehr mit uns als Menschen etwas zu tun als mit den jeweiligen Systemen. Um das nachzuvollziehen, möchte ich Ihnen einen kleinen Rückblick geben – denn die Möglichkeit, »Künstliche Intelligenz« in Form von Computern zu erschaffen, wurde schon in den 1950ern diskutiert, genauso wie die Frage, wie intelligent sie eigentlich werden können.

KAPITEL 3

Künstliche Intelligenz – ein schillerndes Versprechen aus den 1950ern

Die Idee, Maschinen das Zuhören und Sprechen, auch in verschiedenen Sprachen, beizubringen, wurde schon 1950 im berühmten Dartmouth-Forschungsförderantrag zur Künstlichen Intelligenz (im Englischen: *artificial intelligence*, kurz AI) geäußert. Tatsächlich prägte dieser Antrag den Begriff *Künstliche Intelligenz* (kurz KI) überhaupt erst. Seine Autoren waren davon überzeugt, »dass jeder Aspekt des Lernens oder jede andere Eigenschaft der Intelligenz im Prinzip so präzise beschrieben werden kann, dass eine Maschine sie simulieren kann«.[12] Sie nannten unter anderem den Gebrauch von Sprache unter den Problemen, die es zu lösen galt, und gaben an: »Wir denken, dass ein bedeutender Fortschritt bei einem oder mehreren dieser Probleme erzielt werden kann, wenn eine sorgfältig ausgewählte Gruppe von Wissenschaftlern während eines Sommers gemeinsam daran arbeitet.« Eine genaue Definition dessen, was sie unter »Künstlicher Intelligenz« verstehen, haben sie in ihrem Antrag nicht gegeben, aber ich mag die Beschreibung von Marvin Minsky: »KI (ist) die Wissenschaft über das Ertüchtigen von Maschinen, damit sie Dinge tun, die Intelligenz erforderten, wenn ein Mensch sie täte.«[13]

Das Wort »Definition« ist hier eigentlich zu hoch gegriffen, da das Wort ja eigentlich eine scharfe Abgrenzung (von lateinisch *finis*, die Grenze) bezeichnet. In diese Beschreibung würden aber auch moderne Taschenrechner passen – denn sicherlich ist die Kunst beispielsweise der Differentialanalyse von Funktionen etwas, das wir

Abbildung 3: Definition für Künstliche Intelligenz nach Marvin Minsky, einem Pionier auf diesem Feld. Unten links sehen Sie KAI, meinen Roboterfreund, den ich für mein erstes Buch entwickelt habe und der seitdem mietfrei in meinem Kopf lebt. Er versucht, die Menschen um sich herum zu verstehen. In seinem Arbeitszeugnis würde stehen: »Er war stets bemüht!«

normalerweise mit menschlicher Intelligenz verbinden. Aber die meisten Menschen würden heute nicht von Taschenrechnern als KI sprechen. Tatsächlich ist es so, dass Menschen dazu neigen, immer nur das als Künstliche Intelligenz zu bezeichnen, von dem sie bisher dachten, es sei maschinell nicht umzusetzen – was wir als Künstliche Intelligenz bezeichnen, ist daher ein bewegliches Ziel. Taschenrechnerfunktionen sind geradezu das Paradebeispiel dafür, dass eine Tätigkeit, die man für eine intellektuelle hielt, jetzt als eine mechanische wahrgenommen wird. In ähnlicher Weise werden Schachcomputer heute nicht mehr als KI bezeichnet, obwohl sie schon 1997 besser waren als Schachweltmeister. Auch die Spracherkennung von Siri, Alexa oder auf dem Handy nehmen die meisten vermutlich nicht als KI wahr.

Und überraschen Sie perfekte Übersetzungen noch oder ist es für Sie schon zum Alltag geworden, im Spanienurlaub alles schnell übersetzen zu können?

Wussten Sie, dass beispielsweise das Diktieren von Nachrichten ins Smartphone ein KI-System benötigt, genauso wie das Entsperren des Handys per Gesicht oder Fingerabdruck? Die dafür notwendige Technologie ist so alltäglich geworden, dass sie kaum noch wahrgenommen wird – schon gar nicht als Künstliche Intelligenz.

Aber so weit war man in den 1950ern noch lange nicht. In den ersten Versuchen, den Computer »intelligent« zu machen, wurden sogenannte Expertensysteme gebaut: Dazu befragte man Experten und Expertinnen, nach welchen Regeln sie handeln, und baute daraus riesige Mengen von Entscheidungsregeln – z. B. für die Diagnose von Krankheiten oder für das Übersetzen von Texten. Das Problem mit diesen Regelsystemen bestand darin, dass sie einfach zu groß wurden: Das lässt sich anhand der damaligen Übersetzungsprogramme gut demonstrieren. Raten Sie mal, wie eine Maschine aus den 1980ern den folgenden Satz ins Englische übertrug:

»Das passt mir überhaupt nicht in den Kram.«

Ganz klar, d.h. im Englischen:

»That does not usually suit me in the stuff.«

Super, oder? Erinnert mich an einen lieben Freund, der in einer Englischklausur über die »post-war situation in Great Britain« verwirrt war. Er wusste nichts über einen Postkrieg in Großbritannien! But, well, in einem Land, in dem es einen Rosenkrieg gab, ist sicher auch ein Postkrieg möglich! Dass es um die Nach-kriegs-zeit ging, hat er erst bei der Rückgabe der Klausur verstanden …

Solche Übersetzungssysteme enthielten in den 1980ern gern 50 000 bis 80 000 Regeln. Das Problem bei einer falschen Übersetzung besteht nun darin, herauszufinden, welche dieser Regeln verändert oder neu dazugegeben werden müssen, damit die Maschine es beim nächsten Mal richtig macht. Und das wird für den Menschen irgendwann sehr unübersichtlich. Es wurde in den 1980ern also immer offensichtlicher, dass die Welt da draußen viel zu komplex ist, um sie Computern im Detail zu beschreiben. Dasselbe gilt auch für die Kindererziehung. Sie haben sicherlich alle schon einmal versucht, Kindern irgendetwas zu verbieten, was diese dringend wollen. Der Klassiker ist: »Schatz, bitte keine Süßigkeiten vor dem Mittagessen!« Sieht doch nach einer völlig eindeutigen Regel aus! Und wenig später findet man das Kind trotzdem glückselig in seinem Kinderzimmer mit schokoladeverschmiertem Mund. »Mama, das war SCHOKOLADE und keine Süßigkeit!«, behauptet das Kind stolz auf sich selbst, eine so evidente Lücke in Mamas Regel gefunden zu haben. Die Mama kontert und erweitert die Regel mit dem Zusatz: »Schokolade ist auch eine Süßigkeit!« – und das Kind findet die nächste Ausnahme von der Regel. Mein Sohn hat es schon einmal versucht mit: »Ich hatte erst Müsli und dann noch ein Brötchen – das war ja schon mein Mittagessen!« Also musste ich auch noch das Wort Mittagessen definieren! Fazit: Wir können weder Computern noch Kindern die Welt in all ihren Details durch starre Regeln beschreiben, ohne dass wir uns selbst in diesem Regelwerk verlieren und damit bei Reparaturbedürfnissen hilflos vor der Komplexität der Regeln versagen.

Die Idee, dass »jeder Aspekt des Lernens oder jede andere Eigenschaft der Intelligenz im Prinzip so präzise beschrieben werden kann, dass eine Maschine sie simulieren kann«, wie McCarthy und Co-Autoren es in ihrem Forschungsförderantrag formulierten, musste in den 1980ern beerdigt werden – zumindest dann, wenn wir Menschen den Maschinen die Welt präzise beschreiben sollen. Allgemein gilt, dass Systeme mit einer Menge von statischen (unveränderlichen) Regeln scheitern, wenn sie mit der Welt in all ihrer Komplexität interagieren sollen.

Schon in den 1980ern gab es eine Alternative zu diesem Vorgehen: das sogenannte »maschinelle Lernen«, das es heute ermöglicht, dass Autos weitgehend autonom fahren, dass Sie Ihr Handy mit Ihrem Gesicht oder Fingerabdruck entsperren können und dass Bilder, Texte oder Videos basierend auf kurzen Beschreibungen generiert werden können.