IN DIESEM KAPITEL

- Wie Künstliche Intelligenz technisch funktioniert und wie sie definiert wird
- Was es mit der sogenannten Halluzination auf sich hat und was daran gefährlich ist
- Welche bahnbrechenden historischen Entwicklungen notwendig waren
- Eine Erläuterung der KI-Modelle und ihrer Einsatzzwecke

Kapitel 1
Generieren mit KI

Bevor wir uns mit praktischen Einsatzzwecken Künstlicher Intelligenz beschäftigen, sollten wir vorab klären, was das ist, wie sie funktioniert und welche Begriffe und Technologien verwendet werden. Doch keine Sorge: Die Theorie wird nicht grau und dröge, sondern so anschaulich wie möglich.

Grundsätzliche Funktion – Blick in den Maschinenraum

Wenn Sie sich noch überhaupt nicht mit Künstlicher Intelligenz beschäftigt haben (und das ist völlig in Ordnung, denn Sie haben sich das Buch gekauft), stellen Sie sich vielleicht die Frage: Was ist KI und wofür kann ich sie nutzen?

Das Offensichtlichste vorneweg:

- ✔ *KI*, also *Künstliche Intelligenz*, im Englischen *AI*, also *Artificial Intelligence* (artifizielle Intelligenz), ist ein Teilbereich der Informatik und umfasst alle Methoden, die es Maschinen ermöglicht, intelligentes Verhalten zu zeigen.
- ✔ Forschung im KI-Bereich gibt es schon relativ lange. Da jedoch der Begriff *Intelligenz* an sich nur schwammig definiert ist, wird es für die künstliche Variante nicht viel konkreter. Insofern kann KI sehr viel oder auch sehr wenig bedeuten und wird manchmal aus Marketing-Gründen als sehr dehnbarer Begriff missbraucht.

✔ *Generative KI* (*GenAI*) nennt man es, wenn sie (völlig) neue Inhalte wie Texte, Töne oder Bilder erzeugt und in natürlicher Sprache gesteuert werden kann. Sie verwendet in Modellen »gespeicherte« Trainingsdaten, um neue Probleme individuell zu lösen. Um diese geht es in diesem Buch besonders.

Die Sache mit der Statistik: Mit Vorschlägen schneller ans Ziel

Vermutlich haben Sie heute bereits KI-Technik verwendet, ohne es gewusst zu haben. Als Sie beispielsweise jemandem auf eine Textnachricht geantwortet haben: Ihr Smartphone hat Textbausteine vorgesch… ~~vorgeschrieben~~ vorgeschlagen, und Sie haben das passende ausgewählt oder vielmehr gesagt: »Passt schon«. So oder so ähnlich läuft es. Was da arbeitet, ist mindestens eine Vorstufe der Künstlichen Intelligenz.

KI funktioniert mit Wahrscheinlichkeitsrechnung. Sie hat Muster gelernt und folgt Regelwerken und Statistik, um passende Wörter, Bilder oder Töne zu präsentieren. Dank Wortvorschlägen kommen Sie auf dem Weg zu Ihrer Textnachricht schneller ans Ziel, wie *Abbildung 1.1* zeigt. Das nächste passende Wort wird dort schon vorgeschlagen und muss nur angetippt werden. Manchmal trifft das vorgeschlagene Wort haargenau ihre Vorstellung, aber selbst wenn nicht, sind doch oft genug so naheliegende Vorschläge dabei, die akzeptiert werden können, ohne den gewünschten Inhalt komplett zu verfälschen. Deswegen schlägt die App auch gleich mehrere auf einmal vor, und eines wird meistens passen. Hauptsache, es geht schnell, denn wer will schon übermäßig viel Zeit mit einer Textnachricht verlieren?

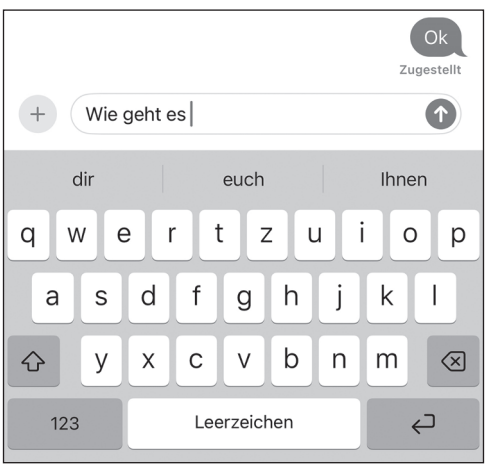

Abbildung 1.1: Beispiel für KI-Technologie – Die Textvervollständigung schlägt mögliche Wörter »dir/euch/Ihnen« vor (Apple QuickType in iMessage am iPhone)

Wenn Ihr E-Mail-Postfach mal aufgeräumt und virusfrei ist, liegt das an dieser Technologie: Filterlisten von Wörtern, die häufig in Spam-Nachrichten vorkommen, sorgen dafür, dass weniger Unliebsames durchdringt. Als Spam markieren trainiert diese Listen. Spamfilter mit Worterkennung und Diktier- oder Übersetzungswerkzeuge sind im Kern nicht viel anders als KI-Technik oder eine automatisierte Vorstufe davon, das *Maschinelle Lernen*.

Von Big Data zum KI-Modell

Wie genau läuft das eigentlich ab? Überall, wo viele Daten, egal ob strukturiert oder zusammenhanglos, vorhanden sind, können KI-Systeme ihre Fähigkeiten entfalten. Anders als bei Software-Entwicklung, die dem *Wenn-Dann-Schema* folgt und die alle potenziellen Möglichkeiten vorgibt, soll der Computer beim *Maschinellen Lernen* selbstständig aus den Daten Schlüsse ziehen.

Beispiele für Ergebnisse von Maschinellem Lernen

- ✔ **Vorhersagen** auf Basis vorliegender Daten: Unternehmensumsatz, Energieverbrauch eines Gerätes, Verkehrsfluss, Wetter
- ✔ **Wahrscheinlichkeiten** berechnen: Spamwahrscheinlichkeit, Kaufeinscheidung, Kündigungswahrscheinlichkeit
- ✔ **Zusammenhänge und Gruppen** erkennen: Wölfe und Hunde sind verwandt, Mops und Dalmatiner sind Hunderassen, Alpakas sind keine Lamas
- ✔ **Kürzen und Verknappen** von Inhalten: Texte ohne größeren Inhaltsverlust zusammenfassen, verlustfrei komprimieren

Füttert man also ein Computersystem mit ausreichend vielen Daten zu Kaufentscheidungen von Kunden und trainiert es somit, kann es berechnen, wie wahrscheinlich der Kauf eines anderen Produktes in der Zukunft stattfinden wird. Weil dabei unglaublich viele Daten einfließen, spricht man von *Big Data*. Nur mit vielen Daten kann Maschinelles Lernen funktionieren.

Doch nur viele Daten allein machen noch keine Künstliche Intelligenz aus. *Kunden, die das eine Produkt kauften, kauften auch …* – damit ist Online-Händler Amazon bekannt geworden. Wenn Sie an anderer Stelle die sechste Käsereibe vorgeschlagen bekommen, obwohl Sie schon zwei besitzen, ist das System dahinter nicht gerade intelligent.

So funktioniert Maschinelles Lernen (Machine Learning): Um der Maschine möglichst viel zum Lernen zu geben, sind gigantische Trainingsdaten erforderlich: *Big Data*. Das können strukturierte oder unstrukturierte Daten sein, die mit relativ geringem Aufwand von der Computerhardware analysiert und dem Lernalgorithmus zugeführt werden.

Es genügt nicht, Unmengen von Daten in ein Modell einzufügen. Die Daten müssen für den jeweiligen Zweck ausgewählt, teils aufwendig vorbereitet und bereinigt werden, damit daraus gelernt werden kann:

- ✔ Zum Training werden Trainingsdaten aufbereitet und der KI präsentiert.
- ✔ Praktisch alle relevanten KI-Technologien benutzen *Künstliche Neuronale Netze*, die dem Aufbau eines menschlichen Gehirns im Sinne von über Synapsen verbundene Neuronen grob nachempfunden sind.
- ✔ Ähnlich wie in unserem menschlichen Nervensystem werden dabei Signale von der Außenwelt (*Input*) über untereinander verknüpfte und unterschiedlich gewichtete Verbindungen weitergeleitet und in den Neuronen verrechnet.
- ✔ In diesen Verbindungen und Gewichtungen des Neuronalen Netzes ist letztlich das Wissen und damit die Künstliche Intelligenz gespeichert. Dieser Vorgang heißt *Modelltraining* und das Ergebnis das *KI-Modell*.

Tatsächlich wurde lange Zeit das *Maschinelle Lernen* als Fähigkeit von Computersystemen angepriesen, bevor das Produktmarketing den Begriff *Künstliche Intelligenz* in den Vordergrund gestellt hat. Wundern Sie sich also nicht, wenn diese Begriffe teilweise verschwimmen oder synonym behandelt werden. In vielen Fällen finden wir hier alten Wein in neuen Schläuchen.

Der Begriff Künstliche Intelligenz (alles Marketing?)

Über Künstliche Intelligenz gibt es sehr weit verbreitete Irrtümer. Es ist nützlich, sie zu kennen, um auch abzuschätzen, was die Maschine im Stande ist zu leisten und was eben nicht. Und wenn Dinge nicht so laufen, wie sie sollen, möchten Sie vielleicht wissen, wieso das so geschehen ist. Dafür nun ein paar grundlegende Antworten.

Wie intelligent ist Künstliche Intelligenz eigentlich?

Zunächst möchte ich festhalten, dass schon der Begriff *Künstliche Intelligenz* leider oft unreflektiert gebraucht wird. Zwar gelingt es einigen Modellen, einen IQ-Test mit überdurchschnittlichen 120 abzuschließen und einen *Turing-Test* (den Beweis für menschenähnliche Kommunikation) zu bestehen. Der Begriff wird allerdings inflationär benutzt und erfüllt nicht immer die Maßgabe des Erfinders Alan Turing. Klar ist, dass man den Maschinen die Fähigkeit, *Intelligenz* zu entwickeln, nicht mehr aberkennen kann. Allerdings unterscheidet sich die Wahrnehmungsqualität heutiger Modelle wahrscheinlich stark von uns Menschen. Zumindest gibt ihre Architektur keinen Anlass dazu anzunehmen, dass Maschinen Gefühle empfinden können oder gar den Antrieb haben könnten, eigenständig zu handeln. Wenn Sie also die Hoffnung hatten, endlich von einer Maschine verstanden zu werden, muss ich Sie enttäuschen: Zumindest die KI versteht uns Menschen nämlich überhaupt nicht.

Trotzdem wird sie gerne vermenschlicht (diese Vermenschlichung nennt man **Anthropomorphismus**). In der Branche ist es umstritten, ob der KI ein Bewusstsein oder gar Gefühle zuerkannt werden sollen. Einige Menschen bedanken sich sogar bei dem Computer für besonders hilfreiche Antworten, ergänzen Prompt-Befehle mit einem `bitte`, wie sie es bei Mitarbeitenden tun würden, und übertragen auf die Software damit menschliche Eigenschaften.

Sollte es einen KI-Knigge geben? Ist die Vermenschlichung ein Problem? Erst einmal mag sie Ihnen nicht schlimm vorkommen, und auch ich neige gelegentlich dazu, den Chatbot derartig zu loben. Doch langfristig betrachtet engt so ein Verhalten unser Verhältnis zu dieser Technik möglicherweise zu sehr ein. Sie ist ein Werkzeug und kein humanoider Freund und sollte das auch bleiben, damit wir instinktiv verstehen, dass wir uns nicht auf sie verlassen können. Eine Selbstkonditionierung oder -regulation, wenn Sie so wollen, muss da ablaufen. Hier sind es *Künstliche Neuronale Netzwerke*, die, ähnlich wie in unserem Gehirn, Informationen möglichst sinnvoll miteinander verknüpfen und wieder abrufbar anordnen. Ob die Maschine dabei in menschlicher Weise versteht, was gemeint ist, ist umstritten. KI hat vermutlich kein Bewusstsein oder Bewusstseinszustände. Einige Eigenschaften, die bereits heute

der KI zugeschrieben werden, müssen erst noch erforscht werden und sind bis dahin schlicht **Marketingbegriffe**: *Künstliche Intelligenz* ist nicht immer intelligent, so wie *Cloud Computing* auch nicht Ihre Daten in irgendwelche Wolken am Himmel auslagert, sondern auf anderer Leute Server.

Wie funktioniert der Chatbot?

Wenn wir von *Chatbots* wie bei ChatGPT sprechen, meinen wir eine Software, die darauf trainiert wurde, eine Anweisung von uns Menschen aus einem sogenannten *Prompt* zu verarbeiten und eine darauf passende Antwort auszugeben. Für Sie erscheint das am Ende so, als würden Sie mit einer Freundin sprechen und sie um etwas bitten, doch es schreibt der Computer.

- ✔ Ein **Prompt**, also die Anfrage an die KI, wird in natürlicher Sprache verfasst.
- ✔ Sie haben Zugriff auf eine Fülle und Tiefe von Informationen.
- ✔ Alles geht ohne Programmierkenntnisse in wörtlicher Rede, die einem menschlichen Gespräch in einem Chatprogramm gleicht.
- ✔ Im Gegensatz zur Internetsuche funktioniert der Chatbot über Dialoge.
- ✔ Die Ausgabe der KI ist auch nicht nur auf Text beschränkt. Sie kann gleich in verschiedenen Stimmen ausgegeben werden, Musik oder Bilder erzeugen oder ein angebundenes Programm steuern. Das nennt man dann *multimodal*.

In den Dialog mit dem Chatbot treten

Wenn Sie Google oder eine andere Suchmaschine benutzen und das Ergebnis verfeinern möchten, fangen Sie jedes Mal von vorne an. Bei einem Chatbot können Sie innerhalb ihrer Anfrage immer wieder Bezug nehmen auf frühere Fragen oder Antworten. Ganz so, wie die Freundin sich ja auch an frühere Gespräche erinnert und darauf Bezug nehmen kann. Aber die KI ist mehr als eine dankbare Gesprächspartnerin oder eine bessere Internetsuche. Sie liefert Ihnen die Ergebnisse in Form einer direkten Antwort. Sie bekommen also nicht nur möglichst passende Suchergebnisse als Auflistung von Internetseiten präsentiert, die Sie mühsam durchklicken müssen.

Dem Chatbot eine Rolle zuweisen

Die Künstliche Intelligenz ist je nach Umfang auch dazu in der Lage, die Rolle einer anderen (realen oder fiktiven) Person einzunehmen und ihre Weltsicht in die Antwort einzubauen. Sie haben, wenn Sie mögen, Albert Einstein oder Kleopatra vor sich und können sie alles fragen. Bei Bedarf sogar in Mundart, gereimt oder gar mit ironischem Unterton. Auch einen Kunden oder Zuschauenden von Ihnen kann die KI spielen und Ihr Produkt analysieren (Näheres dazu erfahren Sie in *Kapitel 8* und Beispiele finden Sie in *Kapitel 9*).

Doch Sie sollten dem Chatbot nicht ungeprüft alles glauben, was er schreibt.

Halluzination – was Dreijährige und die KI gemeinsam haben

Es könnte so schön sein: Wir fragen die Künstliche Intelligenz und erhalten eine Antwort oder Einschätzung, die perfekt und zu 100 Prozent auf Basis vorliegender Fakten stimmt. So sieht es zunächst aus. Aber was ist schon perfekt?

Leider hat KI eine unangenehme Angewohnheit, die als ***Halluzinieren*** oder *Halluzination* bezeichnet wird: die Wahrnehmung nicht existenter Dinge oder falscher Tatsachen. Die Wortvorschläge sind nämlich fehleranfällig. Sind die Antworten nicht ausreichend reflektiert, produziert die KI, statt die Wahrheit zu erzählen, einfach sehr glaubwürdigen Quatsch, der nicht belegt ist und absolut nicht stimmt. Das kennen Sie sicherlich von einigen Menschen (hier bekannt als *Kauderwelsch*, *Konfabulieren* oder *Bullshit*). Im Grunde kann man wohl ein gesundes Maß an Halluzination mit **Fantasie** bei Menschen gleichsetzen, die gelernte Dinge sinnvoll weiterentwickelt. Zu viel davon, oder wenn es um Faktentreue geht, ist jedoch schädlich.

Prominentester Fall von Halluzination (bisher)

Wie jede andere Software macht auch diese Fehler. Diese schmerzliche Erfahrung muss Michael Cohen, Ex-Anwalt von US-Präsident Donald Trump, machen, als er Ende 2023 im Rahmen eines Gerichtsprozesses nach ähnlichen Fällen wie seinem sucht, um diese dem Gericht vorzulegen und damit seinen Argumenten Nachdruck zu verleihen. Dazu übernimmt er leichtfertig die vorgeschlagenen Zitate aus angeblichen Gerichtsakten von der KI *Google Bard* (jetzt *Gemini*). Aber die Sache fliegt auf und wird sehr peinlich für den Anwalt: Schnell stellt sich nämlich heraus, dass die **Zitate von der KI ausgedacht** sind. So wird die Suche nach einem Präzedenzfall selbst zu einem. Es bleibt nicht der letzte, wie Sie in *Kapitel 21* noch lesen werden. Was ist hier passiert?

So kann es zu Halluzination kommen

Im Training bekommt KI beigebracht, Fragen zufriedenstellend zu beantworten. Das versucht sie in einigen Fällen selbst dann, wenn sie gar keine passende Antwort hat. Praktisch jedes Ergebnis, das nicht durch die Trainingsdaten gerechtfertigt ist, kann als Halluzination bezeichnet werden. Technisch kann es viele mögliche Gründe für dieses Phänomen geben:

- ✔ KI wurde mit unausgeglichenen Quellen trainiert (siehe *Bias* in *Kapitel 21*).
- ✔ KI wurde mit irreführenden Quellen trainiert und priorisiert sie falsch (zum Beispiel Lügen aus einem Online-Forum, die als wahre Inhalte ausgegeben werden).
- ✔ KI »interpretiert« Trainingsdaten falsch (zum Beispiel Scherze, Ironie, Sarkasmus).
- ✔ KI wurde falsch trainiert (absichtlich oder unabsichtlich) oder von Cyberkriminellen mit manipulativen Befehlen, der sogenannten *Prompt Injection,* absichtlich beeinflusst.

- ✔ In Ausnahmefällen kann auch ein Missverständnis zwischen Menschen und KI zu vermuteter Halluzination führen, die in Wahrheit gar keine ist. Beispielsweise wenn der Mensch die von der KI erkannten Zusammenhänge in den Daten nicht nachvollziehen kann, wenn er sie mit falschen oder doppeldeutigen Fragen zu unsinnigen Antworten verleitet oder die KI versucht, witzig oder ironisch zu sein.

Vereinfacht gesagt, ist *Halluzination* bei generativer Künstlicher Intelligenz eine oft überzeugend vorgetragene falsche Antwort. Dabei können sogar erfundene oder aus dem Zusammenhang gerissene Quellenangaben vorgebracht werden. Das psychologische Konstrukt des Halluzinierens wird hier auf die Technologie übertragen. Ihr Entstehen zu verstehen und zu vermeiden, gehört zu den größten Herausforderungen im Umgang mit KI.

Die KI ist wie ein fantasierendes Kleinkind im Redefluss

Die Qualität der Antworten hängt von vielen Faktoren ab und kann beizeiten komplett danebenliegen. Aber wieso ist das so? Um es nicht unnötig kompliziert zu machen, vergleiche ich das mal mit meinem (zum Zeitpunkt, als ich dieses Kapitel schreibe) dreijährigen Sohn:

Selbst wenn Sie keine Kinder haben, werden Sie mitbekommen haben, mit welcher **blühenden Fantasie und Vorstellungskraft** die Kleinen ihren Alltag beschreiben können. Wirft man ihnen im Gespräch ein Stichwort zu, verknüpfen Sie es mit irgendwann schon erlebten, gehörten und vielleicht ähnlich klingenden Dingen. Träume und Realität können Kinder vor dem fünften Geburtstag nicht immer trennscharf auseinanderhalten. Eine Zeit lang glauben sie fest an Magie, Fabelwesen oder Wunder (oder dass sich ihr Kinderzimmer von selbst aufräumt, während sie in der Kita sind). Da Dreijährige meist noch keinen zeitlichen Kontext herstellen können, ist alles, was in der Vergangenheit stattgefunden hat, »gestern« oder »vorhin« geschehen: Gestern hatte ich Geburtstag, gestern war Weihnachten, gestern war ich am Spielplatz.

Sie sind stolz, wenn sie schon sprechen können und einmal zu Wort kommen: **Der Redefluss soll nicht abbrechen**. Die Kleinen wollen eine Nachfrage irgendwie beantworten und gegenüber Erwachsenen *smart* wirken, selbst wenn sie eigentlich die richtige Antwort oder den Zusammenhang nicht kennen. Also reimen sie sich etwas zusammen. Ähnlich ist es mit der KI, die darauf ausgerichtet ist, besonders überzeugend in möglichst allen Fällen eine Auskunft zu geben und dafür auf jeden Bereich ihrer Trainingsdaten zurückgreift und vieles miteinander vermischt. Selbst wenn sie eigentlich gar keine wirkliche Antwort hat.

»Früher haben die Ritter gegen die Dinosaurier gekämpft …«, reimt sich ein Dreijähriger nach der Lektüre verschiedener Sachbücher glaubhaft die Welt zusammen. So etwas kann der Künstlichen Intelligenz auch passieren.

- ✔ Rechnen Sie in jeder Antwort oder Berechnung der KI mit *Halluzination*. Sie sollten Ihre menschliche Intelligenz im Umgang mit der künstlichen nicht abschalten und ganz genau aufpassen. Auch, wenn es schnell gehen soll.

- ✔ Selbst Rückfragen sind nicht dazu geeignet, diese Fehlinterpretation zu entlarven. Manchmal können andere – im selben Dialog aber häufig die gleichen – Fehlinformationen herauskommen. Da müssen Sie schon an anderer Stelle recherchieren und die richtigen Quellen prüfen. Da inzwischen viele Inhalte im Internet von KI mitgeschrieben wurden, sind Bücher (noch) eine gute Verifikationsmöglichkeit. Oder Sie rufen mal bei Experten an und fragen persönlich.
- ✔ Auf die Künstliche Intelligenz bezogen, können Sie den wertvollen Ratschlag einer Kita-Erzieherin meines Sohnes beherzigen: »Glauben Sie bitte nicht alles, was Ihr Kind zu Hause über uns erzählt. Wir tun es umgekehrt auch nicht.«

An Lösungen gegen Halluzination wird gearbeitet

Mit diesem Phänomen der Halluzination haben praktisch alle KI-Modelle mehr oder weniger zu kämpfen. Aus diesem Grund geben sie auch Warnungen zur Unzuverlässigkeit der Antworten ab. Bei ChatGPT steht beispielsweise unter dem Chatfenster: »ChatGPT kann Fehler machen. Überprüfe wichtige Informationen.« Aber die sind schnell übersehen und vergessen.

Es gibt von Herstellerseite einige Ansätze, dem Halluzinieren vorzubeugen:

- ✔ Bei besonders auffälligen und nachvollziehbaren Fehlinterpretationen versuchen die Unternehmen mal mehr, mal weniger erfolgreich, dagegen anzutrainieren.
- ✔ Andere Ansätze motivieren das KI-Modell zu einer Art Selbstdiagnose. Dabei zerlegt das System die Antworten in einzelne Aussagen, überprüft deren Relevanz und vergleicht diese mit anderen Antworten oder Suchergebnissen im Internet.
- ✔ Das sogenannte *Reasoning*, diese KI-Fähigkeit, Schlüsse zu ziehen, ist beispielsweise von OpenAI in Sprachmodellen wie o1 eingebracht, um verlässlichere Antworten zu liefern. Sie lernen es in *Kapitel 5* näher kennen. Auch das quelloffene Sprachmodell Reflection (siehe *Kapitel 7*) versucht so etwas.

So vielversprechend die Ansätze sind, sie stecken eben noch im Experimentierstadium, wie die ganze KI-Technologie auch (siehe *Kapitel 2*).

Werfen wir doch mal einen Blick in die Modelle hinein, um zu verstehen, wie und wo der Bullshit und das Kauderwelsch so schön konfabulieren können.

Eine kurze Chronologie – was bisher geschah …

Um zu verstehen, wie die Technologie funktioniert und was ihr noch fehlt, hilft ein Blick zurück über die Schulter: Denn lange vor ChatGPT und autonomen Robotern haben sich Entwicklungen abgespielt, die bis heute wirken. Im Kern steht immer das lang ersehnte Ziel, ein Abbild des Menschen herzustellen. Besonders von dem, was sein Denken ausmacht: dem Gehirn.

Der menschliche Traum von künstlicher Assistenz

Die Menschheit träumt seit Jahrtausenden davon, gottgleich neues Leben und autonome Helfer erschaffen zu können. An Beispielen mangelt es nicht.

Historische und literarische Vorbilder

Im sumerischen und babylonischen **Gilgameš-Epos**, der zu den ältesten bekannten schriftlichen Aufzeichnungen überhaupt zählt, oder in der griechischen und römischen Mythologie und früheren Erzählungen aus Ägypten und China wimmelt es vor künstlichen Wesen und Automaten.

Auf das 12. Jahrhundert datiert, finden sich Varianten der jüdischen Erzählung vom ***Golem*** (hebräisch so etwas wie *formlose Masse*; *dummer und unfertiger Mensch*), der Kommandos ausführt, wenn man ihm einen beschrifteten Zettel unter die Zunge legt. Klingt für mich nach Lochkartenprogrammierung.

Goethe beschrieb im **Zauberlehrling** seine Version des außer Kontrolle geratenen Helferleins und wurde seinerseits dreist von Disney kopiert. Leonardo Da Vinci baute erste menschenähnliche **Roboter** (und damit meine ich nicht die Mona Lisa), die er mit Schnüren zum Aufstehen bewegen konnte.

Mechanical Turk: Der Betrug hat Tradition

Nicht immer ist hochentwickelte Technik beteiligt, wenn es danach aussieht. Vielleicht haben Sie vom Begriff ***Schachtürke*** oder ***Mechanical Turk*** gehört.

Jahrmarktattraktion und höfische Unterhaltung

Gemeint ist damit einer der größten Täuschungsfälle der Geschichte: Der österreichisch-ungarische Hofbeamte und Mechaniker Wolfgang von Kempelen ist in Wahrheit ein früher Illusionist. Er präsentiert 1769 seinem Publikum eine Kiste mit einer scheinbar automatisch funktionierenden, traditionell orientalisch gekleideten Puppe, dem »schachspielenden Türken« (*Abbildung 1.2*), der nahezu jedes Schachspiel gewinnen kann. Was die gegnerische Seite und die Zuschauenden nicht wissen: Im Inneren des Apparates verbirgt sich ein kleiner Mensch, der alles steuert (und offensichtlich sehr gut Schach spielen kann). Es gibt gar keine ausgeklügelte und überlegene Maschine.

Der Trick soll erstaunlich lange durchgegangen sein – mindestens ein Jahrzehnt, später geht das Gerät oder Nachbauten davon sogar unter anderen Besitzern noch für über 80 Jahre auf Europa- und Amerikatournee. Vielleicht wollten die Leute glauben, dass es funktioniert.

Heutige Erscheinungsform der »Turker«

In der Computerwelt spricht man heute ebenfalls vom **Mechanical Turk**, wenn bei scheinbar automatischen Prozessen von Menschenhand eingegriffen und nachgeholfen werden muss, ohne dass Nutzende das am Ende erkennen.

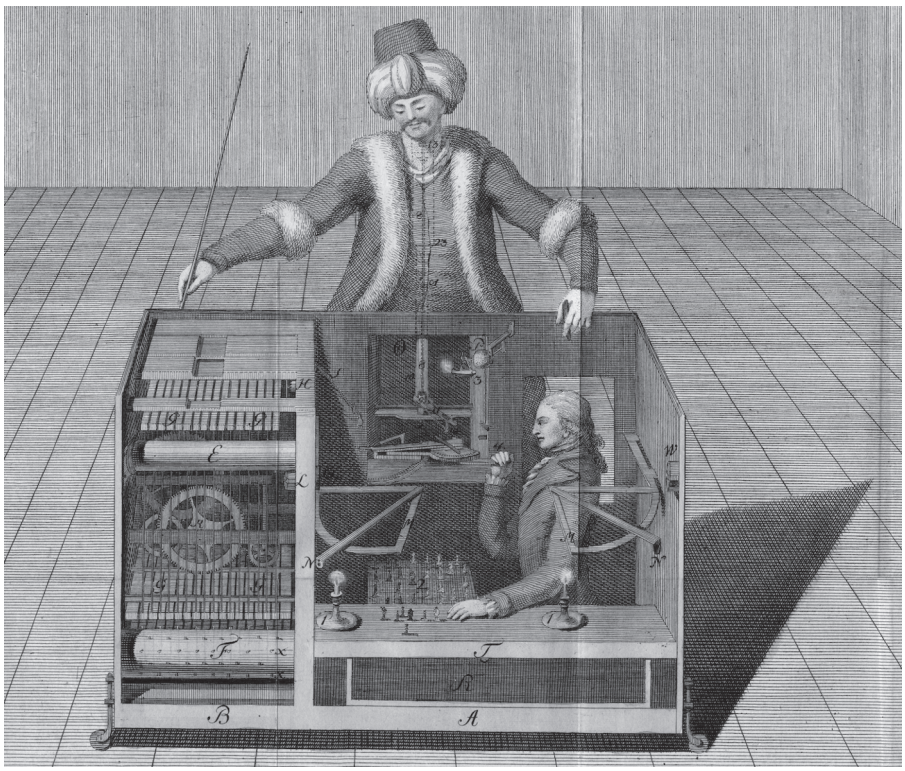

Abbildung 1.2: Mechanical Turk; Kupferstich aus »Ueber den Schachspieler des Herrn von Kempelen und dessen Nachbildung«, Joseph Friedrich zu Racknitz; Breitkopf, 1789

Die Idee vom Mechanical Turk hat sich bis ins 21. Jahrhundert gerettet:

- ✔ **Amazon** beispielsweise bietet unter `MTurk.com` seine *Clickworker* aus aller Welt, die sogenannten »Turker«, an, die für schnell zu erledigende Aufgaben wie der Erkennung von Bildern, dem Transkribieren von Texten oder dem Umstellen von Informationen in Datensätzen wenige US-Cent verdienen können.
- ✔ Wenn Sie dachten, das liefe längst automatisch: Dienstleister **Appen** bewirbt, die KI-Modelle von Google, Meta und anderen Techfirmen mit über einer Million Clickworkern zu optimieren – also zu korrigieren. Wie viel Zeit Niedriglöhner dabei für einen fundierten Faktencheck der Antworten haben, ist nicht bekannt.
- ✔ Elon Musk präsentiert auf einem Tesla-Event 2024 die angeblich autonomen KI-Roboter »Optimus«. Sie schenken Drinks aus und bespaßen Gäste. Schnell stellt sich heraus, dass sie ferngesteuert waren und nicht einmal die Sprachausgabe vom Computer kam. Hätte man ahnen können, denn die Show fand in Hollywood statt.

Nicht überall, wo KI draufsteht, ist auch welche drin. Übertreibung im Marketing schadet der ganzen Branche. Man spricht dabei von *AI-washing*, analog zu *Greenwashing* beim angeblichen Umweltschutz.

KI als Ziel einer langen Reise von Erkenntnissen

Parallel zu Träumereien und vorgetäuschten Revolutionen schreitet bekanntermaßen auch die Wissenschaft voran.

Künstliche Intelligenz als eigene wissenschaftliche Disziplin

Bei all der Vorarbeit ist erstaunlich, dass der eigentliche Titel *Artificial Intelligence* noch gar nicht so alt ist: Der US-Informatiker John McCarthy und seine Nerd-Kollegen möchten mit diesem Kunstbegriff 1955 ihren Förderantrag bei der Rockefeller-Stiftung durchbringen. Zwar erhalten sie am Ende mit 7500 US-Dollar nur etwas mehr als die Hälfte der beantragten Summe, und es kommen von 47 geladenen Wissenschaftlern auch nur gut ein Dutzend. Die **Dartmouth Conference** im Sommer 1956 wird dennoch zur Geburtsstunde der KI, denn die wichtigsten Errungenschaften werden in der Folgezeit von den Konferenzteilnehmern oder deren Studierenden erzielt.

Meilensteine der KI-Entwicklung

Ohne Sie mit zu vielen Details langweilen zu wollen, die mathematischen Grundlagen für KI sind extrem früh gelegt worden.

Meilenstein	Jahr	Bedeutung
Formalisierung, Logik	Etwa 1000 vor Christus	*Denken* wird in China, Indien und im antiken Griechenland in mechanische Prozesse und Einzelteile zerlegt wie Dim Sum, Thali und Gyros.
Deduktion	400–200 vor Christus	Philosoph Aristoteles und Mathematiker Euklid forschen wie wild an logischen Schlussfolgerungen.
Algorithmus	9. Jahrhundert	Muhammad al-Chwarizmi (»Algorismi«) erfindet die *Algebra* zur Problemlösung und schafft spätere Probleme für Generationen von Oberstufenschülern.
Differentialrechnung und Kettenregel	17. Jahrhundert	Keks-Namensgeber Gottfried Wilhelm Leibniz grübelt an einer *vergleichbaren universellen Formensprache* (*characteristica universalis*) und entwickelt die mathematische Basis für das Training der Sprach- und Diffusionsmodelle. Ohne die Kettenregel könnten neuronale Netzwerke nicht lernen.
Computeralgorithmus	1843	Die Britin Ada Lovelace arbeitet mit dem Mathematiker Charles Babbage zusammen. Sie schreibt die weltweit erste Software (per Hand, denn es gibt noch keine Computer) und prophezeit: Diese Technologie wird universell einsetzbar sein!

Meilenstein	Jahr	Bedeutung
Logikkalkül	1847	George Boole entwickelt die mathematische Sprache, in der Computer bis heute »denken«. In der *Booleschen Algebra* werden komplexe Aussagen auf einfache *JA/NEIN-Entscheidungen* reduziert und über die *Booleschen Operatoren UND*, *ODER*, *NICHT* miteinander verknüpft. Sie sind das Fundament für Schaltkreise und die Informatik. Ohne sie gäbe es keine Programmierung.
Turing-Maschine	1936	Mit der Entwicklung des theoretischen Konzepts der *Turing-Maschine* legt der Brite Alan Turing den Grundstein dafür, welche Probleme überhaupt berechenbar sind.
Erster Computer der Welt	1941	Aus 2600 gebrauchten Telefonrelais baut Konrad Zuse in Berlin eine revolutionäre Maschine, die *Z3*, die das binäre System aus Nullen und Einsen nutzt und programmierbar ist. Erst nach dem Zweiten Weltkrieg erfährt die Welt davon. Das Original wird bei einem Bombenangriff zerstört. Ein Nachbau steht im Deutschen Museum in München.
Robotergesetze	1942	Isaac Asimovs Geschichte vom durchdrehenden Roboter mahnt zu vorbeugender Technologieregulierung und wird 60 Jahre später als »I, Robot« mit Will Smith schlecht verfilmt.
McCulloch-Pitts-Neuron	1943	Das erste simple Neuronenmodell, das zeigt, dass *Boolesche Logik* durch Netzwerke berechnet werden kann. Die beiden Forscher legen die Basis heutiger Künstlicher Neuronaler Netze.
Turing-Test	1950	Schon wieder Turing: Alan Turing legt als »Vater der KI« den Grundstein für KI-Technologie. Sein Test definiert ein Kriterium dafür, wann eine Maschine menschenähnlich kommunizieren kann.
Dartmouth Conference	1956	Geburt der KI als Wissenschaft! Der heute legendäre Brainstorming-Workshop vernetzt und motiviert knapp ein Dutzend Wissenschaftler unterschiedlicher Disziplinen, darunter neben John McCarthy auch Marvin Minsky, Claude Shannon und Nathaniel Rochester.
Perzeptron	1957	Mit seinem Neuronenmodell entwickelt Frank Rosenblatt eine erste limitierte Möglichkeit, KI zu trainieren *(perception - wahrnehmen)*.
Erste trainierbare Mehrschichtmodelle	1965	Die Grundlage des heutigen *Deep Learning*: Alexey Ivakhnenko und Valentin Lapa erweitern Rosenblatts Modell, indem sie Neuronen zu einem Netzwerk mit mehreren Schichten kombinieren und die Fähigkeit verleihen, komplexere Entscheidungen zu treffen. Ihre Arbeit wird im Westen erst nach dem Kalten Krieg umfassend bekannt.

Meilenstein	Jahr	Bedeutung
Chatbot ELIZA	1966	Joseph Weizenbaum erfindet die *Mensch-Maschine-Kommunikation* mit Schlüsselwörtern. Weit entfernt von ChatGPT, absolut KI-frei.
1. KI-Winter	1969	Marvin Minsky und Seymour Papert zeigen Grenzen der Entwicklung auf und beweisen, dass Rosenblatts Perzeptron nur sehr einfache Entscheidungen treffen kann. Ihre und eine weitere Veröffentlichung leiten eine Phase der Ernüchterung in der KI-Forschung ein: Projekte werden gestoppt, Gelder gekürzt.
Backpropagation, Rekurrente Neuronale Netzwerke (RNN) und Reinforcement Learning (RL)	1980er	Wendepunkt: Mehrschichtige neuronale Netzwerke können mit dem *Backpropagation*-Algorithmus basierend auf der Kettenregel von Leibniz effizient trainiert werden, Fehler erkennen und sich selbst verbessern. Daran wurde seit den 60ern geforscht. RNNs können zeitabhängige Daten wie Sprache und Musik verarbeiten, indem sie sich vorherige Eingaben merken. Sie vergessen diese jedoch oft schnell wieder. Maschinen lernen außerdem zum ersten Mal selbstständig allein durch Interaktion mit ihrer wahrgenommenen Umgebung (RL).
2. KI-Winter	1987	Nach einer kurzen Erholungsphase, in der große Hoffnungen in Expertensysteme gesetzt wurden, folgte erneut ein sogenannter KI-Winter.
Long Short-Term Memory (LSTM)	1997	Eine neue Architektur, die auf RNNs aufbaut, ermöglicht es, dass Maschinen sich Informationen über längere Zeit merken können. Das wiederum ermöglicht es, komplexe Zusammenhänge in zeitabhängigen Daten wie Sprache oder Musik besser zu verstehen und zu verarbeiten.
IBM Watson	2011	Der Moment medialer Öffentlichkeit: Eine KI kann um die Ecke denken und gewinnt damit die TV-Quizshow Jeopardy. Die 77.147 Dollar Preisgeld werden gespendet.
Convolutional Neural Networks (CNN)	2012	Maschinen können Bilder erkennen und klassifizieren. Das Netzwerk *AlexNet* setzt Maßstäbe und gewinnt den *ImageNet*-Wettbewerb für Deep Learning und Objekterkennung.
Variational Autoencoder (VAE)	2013	Die neue Form Künstlicher Neuronaler Netze kann Bilder, Musik oder 3D-Modelle auf ihre wesentlichen Merkmale reduzieren und daraus neue, ähnliche Inhalte generieren.
Generative Adversarial Networks (GAN)	2014	Auf dem Heimweg aus einer Kneipe erfindet der US-Informatiker Ian Goodfellow den Ansatz eines Wettbewerbs konkurrierender Netzwerke, die sich gegenseitig verbessern und für realistischere Bilder, Videos oder Musik sorgen.

Meilenstein	Jahr	Bedeutung
DeepMinds AlphaGo	2016	Im komplexesten Strategiespiel der Welt, Go, wird der amtierende Weltmeister von einer KI besiegt – eine Demonstration der Leistungsfähigkeit von RL mit Deep Learning. Google kauft den Laden, packt ihn zu den anderen Dingen und vergisst, das Produkt zu veröffentlichen.
Transformer-Architektur (das T in ChatGPT)	2017	*Selbstaufmerksamkeit* als bahnbrechende Idee des Google-Mitarbeiters Ashish Vaswani und Kollegen (»Attention is all you need«): Nicht der Reihe nach, sondern gleichzeitig werden Informationen betrachtet und dabei gewichtet. Jahrzehnte hatte man an solchen Ansätzen geforscht.
Diffusionsmodelle	2020	Eine neue Klasse von Modellen, die durch schrittweise Umkehrung eines Zufallsprozesses besonders realistische Bilder, Videos oder Musik erzeugen können. Sie sind die Grundlage für Anwendungen zur Bildgenerierung wie Midjourney, DALL-E oder Stable Diffusion.
ChatGPT	2022	Im November präsentiert OpenAI *ChatGPT,* und die Welt ist beeindruckt, wo das jetzt auf einmal herkommt.

Tragisches Wunderkind und »KI-Vater«: Alan Turing

Der britische Logiker, Mathematiker, Kryptoanalytiker und Informatiker **Alan Turing** gilt heute als einer der einflussreichsten Theoretiker der Informatik und KI-Lehre. Doch er ist eine ausgesprochen tragische Figur, der viel Unrecht widerfahren ist. Für die erzwungene Hormonbehandlung wegen seiner Homosexualität, die ihn letztlich mit 42 Jahren in den Suizid treibt, entschuldigt sich die britische Regierung erst 2009 posthum.

Seine hohe Begabung soll Turing schon in frühester Kindheit gezeigt haben, als er sich selbst das Lesen beibrachte und sich zu Zahlen und Rätseln hingezogen fühlte. Nach dem Tod seines besten Freundes zieht er sich weiter zurück, wird als schrulliger Theoretiker mit teils autistischen Zügen beschrieben, der auch mal im Frühjahr mit Gasmaske zur Uni radelt, um Heuschnupfen zu entkommen.

- ✔ Nicht nur seine Schriften sind bahnbrechend, er erarbeitet 1936 im Alter von 24 Jahren die ***Turing-Maschine***, die bis heute als Beweis der Berechenbarkeitstheorie herangezogen wird, stellt mit dem ***Turing-Test*** eine Methode zur Erkennbarkeit von Intelligenz auf und konstruiert ab 1945 einen ersten Supercomputer, die ***Automatic Computing Engine***, die weit schneller ist als die Rechenmaschine von **Konrad Zuse** und bis heute Vorbild für den Computerbau ist, letztlich aber **John von Neumann** zugeschrieben wird, der die Idee vermutlich kopiert hat.
- ✔ Im Zweiten Weltkrieg ist Turing einer der wichtigsten Codeknacker der Alliierten und dechiffriert die Enigma-Funkverschlüsselung der Nazis, was den Krieg verkürzt.
- ✔ Nach ihm ist die bedeutendste Informatikauszeichnung **Turing Award** benannt.

Die Captcha-Funktion zur Abwehr von Spam und Bots steht für **C**ompletely **a**utomated **p**ublic **T**uring test to tell **c**omputers and **h**umans **a**part, also *vollautomatischer öffentlicher Turing-Test zur Unterscheidung von Computern und Menschen*. Entgegen dem echten Turing-Test entscheiden hier Computer, ob Sie ein Mensch sind und nicht umgekehrt. Ironie dabei: Mit reCAPTCHA requiriert **Google** Internetsurfende als *Mechanical Turks* und zum KI-Training.

Ein Modell für den Nachbau des Gehirns

Alan Turing sah das menschliche Gehirn als Prototyp für Intelligenz an. Aber wie genau kann unser Nervensystem komplexe Rechenvorgänge ausführen? Wie wirken Synapsen und Neuronen zusammen?

Die Idee von Künstlichen Neuronalen Netzen

Um zu klären, ob und wie das biologische Gehirn Turing-berechenbare Funktionen wirklich berechnen kann, entwickeln der Neurophysiologe **Warren McCulloch** zusammen mit dem Logiker **Walter Pitts** 1943 das ***McCulloch-Pitts-Neuron***, das beweist, dass die Nerventätigkeit im Gehirn streng logischen Gesetzen folgt, nämlich denen von Bools Algebra mit den rund 100 Jahren zuvor formulierten logischen Operatoren und Nullen und Einsen (siehe *Abbildung 1.3*). Sie stellen darin das Zusammenspiel von Nervenzellen im Gehirn extrem vereinfacht und abstrakt dar und machen es berechenbar. So schaffen sie auf basierend von Mathematik und Logik den Baustein, aus dem Künstliche Neuronale Netze bis heute meist basieren. Ohne dieses Modell wären künstliche Intelligenzen völlig undenkbar.

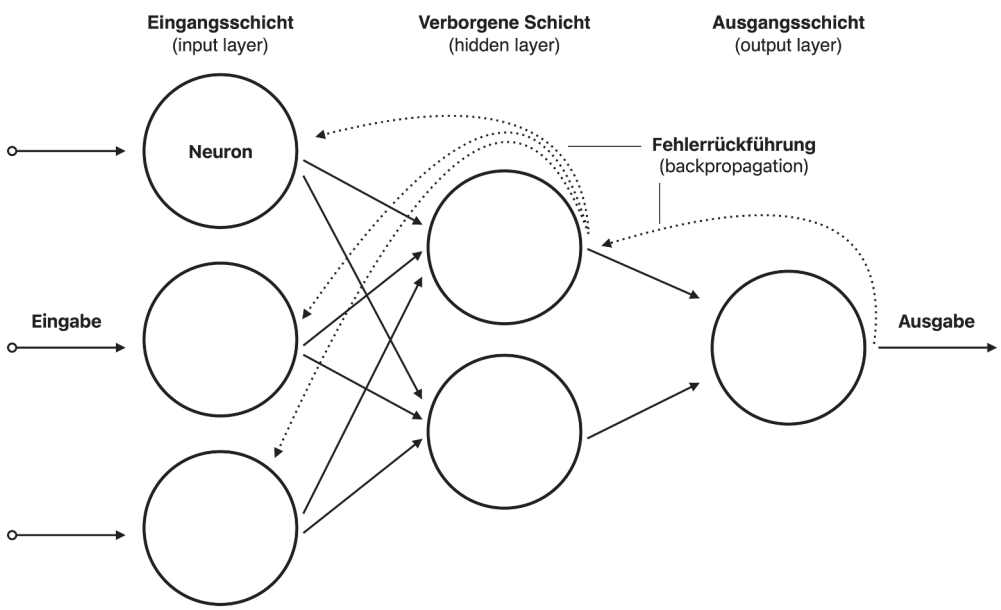

Abbildung 1.3: Netzwerkarchitektur Künstlicher Neuronaler Netze (KNN)

Wie ein **Künstliches Neuronales Netzwerk in der Praxis** funktioniert, können Sie auf `https://playground.tensorflow.org` spielerisch ausprobieren und die Neuronenreaktionen über mehrere Layer nachverfolgen.

Abgrenzung zum menschlichen Gehirn

- ✔ Beim biologischen Gehirn bildet der Cortex (auch als Hirnrinde bekannt) den äußersten Teil und ist in sechs Schichten aufgebaut. Er ist für höhere kognitive Funktionen wie Denken, Planung und Sprache verantwortlich.
- ✔ Künstliche Neuronale Netze (KNN) sind ebenfalls in mehreren Schichten aufgebaut. Diese haben aber keine direkte Entsprechung zu den biologischen Schichten des Cortex.

Ein KI-Gehirn ist kein exakter Nachbau des menschlichen Gehirns aus Kabeln und Prozessoren, sondern existiert virtuell in Computern als Programm.

Künstliche Neuronale Netze imitieren konzeptionell gewisse Abläufe und Strukturen des Gehirns: Die Neuronen werden repräsentiert durch Knoten, die über Kanten miteinander verbunden sind. Diese Verbindungen sind gewichtet, wobei die Gewichtung die Stärke der Verbindung darstellt. Diese Verbindungen entsprechen den Verästelungen der Neuronen im Gehirn und die Gewichtung der Stärke der Synapsen.

Diese Gewichte werden während eines Trainingsprozesses angepasst, bei dem das Netzwerk mit großen Mengen von Daten konfrontiert wird und dadurch lernt, sinnvolle Ausgaben für gegebene Eingaben zu generieren.

Maßgeblicher Unterschied: Wo unser Gehirn durch seine Größe und die Anzahl von Neuronen begrenzt ist, kennt die KI kein Limit. Ein durchschnittlicher Intelligenzquotient (IQ-Wert) von 100, der für Menschen normal ist, wird von der KI bereits übertroffen.

Vergleich des Lernprozesses

Lernprozesse bei KNN unterscheiden sich grundlegend von dem des Gehirns:

- ✔ Das menschliche Gehirn lernt kontinuierlich und kann sich neuen Situationen anpassen.
- ✔ Im Gegensatz dazu fungieren KNN nach dem abgeschlossenen Trainingsprozess wie ein »eingefrorenes« Gehirn: Die angepassten Gewichte bleiben statisch, sodass das KNN nur so flexibel und schlau ist wie ihr letztes Training.

So »lernt« die KI, ein Objekt zu erkennen

Erkennungsprozess

1. Angenommen, Sie präsentieren ein Bild von einem Haustier (*Eingabe*): Neuronen der Eingangsschicht (*Abbildung 1.3*) erkennen vielleicht nur grobe Muster.
2. Diese Information liefert der nächsten Schicht im verborgenen Bereich Hinweise, wo sie nach Armen und Beinen suchen kann.
3. Das wiederum aktiviert diejenigen Neuronen der dritten Schicht, die mit dieser Anordnung etwas anfangen können und Augen identifizieren und so weiter.
4. Die *Abbildung 1.3* können Sie sich vervielfacht vorstellen. In der nächsten Schicht sind Neuronen am stärksten aktiv, die in dieser Anordnung einen Hund sehen.
5. Irgendwann gibt eine letzte Schicht eine Antwort: Am Ende steht fest, dass wir hier ein Tier vor uns haben, und zwar wahrscheinlich einen Hund.

   Eine perfekte Kettenreaktion, die zu einem Ergebnis (*Ausgabe*) führt.

Rückmeldung und Lernen

Wie die KI in intelligenten Mustern denken kann, also diese Informationen richtig gewichtet, lernt sie im Training, wovon es verschiedene Stufen gibt.

- ✔ **Backpropagation** heißt der iterative Prozess, bei dem jeder Fehler eines Neuronalen Netzes von der Ausgabeschicht rückwärts durch alle Schichten *propagiert*, also verbreitet, wird. Es werden nur Muster trainiert. Das ist besonders effektiv. Niemand muss jede einzelne Information einprogrammieren. Mit unterschiedlichen Beispielen wird dieser Prozess so häufig wiederholt, bis das Netz gelernt hat, auf welche Details es für die richtige Antwort ankommt.
- ✔ Ist die ausgegebene Antwort korrekt, kann sie von einem Menschen bestätigt werden (*supervised training*). Falls nicht, werden Verbindungen anders gewichtet, um die KI an die richtige Antwort heranzuführen. Wie eine Lehrerin mit einer Schulklasse bringen von Menschen kuratierte Trainingsdaten der KI bei, ob sie am Ende (*Ausgabe*) richtig oder falsch liegt.
- ✔ Welche Gedanken der KI auf dem Weg dorthin durch den »Kopf« schießen, bleibt unerkannt (*verborgene Schicht*). So, wie jede Schülerin und jeder Schüler individuell denkt und erst anschließend korrigiert wird, wenn die Antwort daneben liegt. Auch bei der KI weiß praktisch niemand, welche Schlüsse sie auf dem Weg zur Entscheidung zieht. Sie wird nur immer und immer wieder trainiert, richtig zu denken, bis nur noch sehr wenig korrigierend eingegriffen werden muss.

Große Modelle und generative KI – Coverversionen und Me-too-Produkte

Sie könnten sich fragen, wieso gerade jetzt diese Entwicklung kommt, dass man mit einem Computer so realistisch chatten und Dinge erschaffen kann, wo es doch schon so lange Künstliche Intelligenz gibt. Dafür sollten wir kurz etwas detaillierter auf die Funktionen von Sprachmodellen schauen und auch ein paar Begrifflichkeiten klären. Dafür wird es in diesem Abschnitt noch einmal sehr technisch. Aber ich denke, damit kommen Sie inzwischen klar, und falls nicht, können Sie ihn jederzeit zum Nachschlagen nutzen.

Geniale Verbindung aus Chatbot und Sprachmodell

Das *Chat* in ChatGPT kommt von der Verbindung zu einem Chatbot.

Ein *Chatbot* ist eine Software, die menschliche Sprache und Konversation simuliert. Nicht alle Chatbots basieren auf Sprachmodellen. Aber Sprachmodelle können als Chatbot auftreten und mit Menschen interagieren.

Chatbots sind keine neue Erfindung. Als Erster erregte *ELIZA* bereits 1966 großes Aufsehen. Das von Joseph Weizenbaum am Massachusetts Institute of Technology (MIT) entwickelte Programm gibt vor, ein echter Psychotherapeut zu sein, und gilt als **Vorreiter der Mensch-Maschine-Kommunikation**.

Bisher funktionierten die meisten Chatbots wie ELIZA mittels musterbasierter Texterkennung, konnten auf Schlüsselwörter reagieren und hatten ansonsten ausweichende Standardfloskeln an Bord, die Unverständnis kaschieren.

Eine Abwandlung von ELIZA, den »digitalen Psychotherapeuten« **Dr. SBAITSO,** können Sie ausprobieren. Damit hatte ich 1992 schon meinen Spaß: `https://archive.org/details/msdos_Dr_Sbaitso_1992`.

Mit KI-Anbindung werden Chatbots zum leicht bedienbaren *Interface*:

- ✔ In früherer Zeit wurden Chatbots nicht von KI, sondern regelbasiert angetrieben. Auf Stichwörter folgt eine Antwort aus vorformulierten Textbausteinen.
- ✔ Mit Künstlicher Intelligenz und dem Zugriff auf große *Sprachmodelle* kann diese Sprachausgabe nun in jeder Situation *live* generiert werden, obwohl niemand die Antwortmöglichkeiten vorhergesehen oder gar vorformuliert hat.

Wie große Sprachmodelle (Large Language Models) arbeiten

Es gibt viele Faktoren, die für den Durchbruch von generativer Künstlicher Intelligenz in unserer Zeit eine Rolle spielen: Neben den Ressourcen für Energie und der Rechenleistung durch Cloud Computing, die früher nicht in dem Ausmaß vorhanden waren, ist es besonders die Tatsache, dass ChatGPT und alle Nachahmerprodukte so einfach zu bedienen sind.

Einfache Bedienung als entscheidender Faktor

Eine einfache Sprache statt Programmierkenntnisse, ein Dialog statt eines langen Befehls. Kurz gesagt: Bisher kam einfach niemand auf die Idee, die beiden Bereiche Sprachmodell und Chatbot miteinander zu verbinden.

Welche Sprachmodelle man dabei verknüpft, beeinflusst die Ausgabe maßgeblich. Es gibt Sprachmodelle, die besonders eloquent sind, oder solche, die besonderes Fachwissen in spezifischen Bereichen haben.

Als **Large Language Models (LLM)**, also besonders große Sprachmodelle, bezeichnet man leistungsstarke Strukturen, die menschliche Sprache verstehen und generieren können. Das sind erst einmal *Künstliche Neuronale Netzwerke*, die kognitiven Fähigkeiten des Gehirns nachahmen, aber mit unfassbar vielen miteinander verbundenen Neuronen und Parametern (large!). Diese können Sie sich erst einmal wie eine große Excel-Tabelle voller Zahlen und Zusammenhänge vorstellen, denn sie arbeiten ausschließlich mit Zahlen: Zahlen werden hineingeschickt, und andere Zahlen kommen hinten heraus. Doch sie gehen über tabellarische Strukturen hinaus und nutzen mehrdimensionale Vektordarstellungen, die semantische Beziehungen zwischen Wörtern abbilden. Da sie jede Form von Inhalt (Text, Bilder, Audio, Video) in Zahlen darstellen können, sind sie so vielseitig.

- ✔ Künstliche Intelligenz hängt inzwischen von großen **Sprachmodellen** ab, in denen das Trainingsmaterial steckt und miteinander verknüpft ist.
- ✔ Die Auswahl des Sprachmodells entscheidet über die Qualität der Ausgabe.
- ✔ Der niedrigschwellige Zugang über Chatbots ist ein Hauptgrund für die rasende Verbreitung und Verknüpfung mit einer ganzen Reihe anderer Technologien.

Begriff der großen Modelle

Auch wenn umgangssprachlich meistens von *Sprachmodellen* die Rede ist und das nicht falsch ist, gibt es mit diesem Begriff inzwischen Schwierigkeiten. Er reicht nicht aus, um den Trainingsumfang und die Dimension zu beschreiben und zu unterscheiden. Hinzu kommt, dass viele Modelle *multimodal* sind.

Das ist nicht nur für Nutzende, sondern auch bei der Gesetzgebung relevant *(siehe Teil V)*. Ein Modell, das Wahlen beeinflussen könnte, soll schließlich stärker reguliert werden als eins, das nur Katzenbilder generiert.

Große Sprachmodelle mit generalisierenden Fähigkeiten nennt man ***Foundation Models*** (zu Deutsch **Grundmodelle** oder **Basismodelle**). Foundation Models sind beispielsweise Googles Gemini, Metas Llama oder OpenAIs ChatGPT. Hier wäre GPT das Sprachmodell, oft wird zur besseren Vergleichbarkeit noch die Version angehängt, also zum Beispiel *GPT-4*.

Lernen große Sprachmodelle wie kleine Kinder?

Was im Sprachmodell exakt abläuft, ist tatsächlich noch gar nicht vollständig erforscht. Grob weiß man, das haben Sie am Anfang dieses Kapitels erfahren, dass das Sprachmodell ein Vorhersagespiel spielt und das nächste passende Wort rät. Doch wie kommen alle diese Wörter zustande?

Menschliches Lernen

Veranschaulichen wir uns im Vergleich dazu einmal, wie Babys sprechen lernen: Niemand drückt ihnen eine Anleitung wie »*Sprechen lernen für Baby-Dummies*« in die Hand. Sie können es irgendwann einfach (oder müssen zur Logopädie). Zunächst hören Kleinkinder, wie jemand etwas sagt. Das sind meistens die Eltern. Irgendwann imitieren sie das, plappern lautmalerisch einzelne Wörter nach und formen erste Sätze. Heißt »*Dada*« jetzt *Papa* oder *Bagger*? Das müssen Eltern zunächst aus dem Kontext erraten. Immer wieder werden diese Kleinen (hoffentlich) anständig korrigiert und lernen die korrekte Aussprache, irreguläre Formen oder Grammatik. Gerade bei fremdsprachigen Liedern kann es vorkommen, dass der Dreijährige sich Lieder wie »laikto-mufit« (*I like to move it*) wünscht, bis ihm jemand erklärt, wie das wirklich heißt.

In einzelnen Bereichen kennt sich das Kind dann bald schon aus und kann sehr viel dazu erzählen, beispielsweise wenn ein Gespräch auf die Feuerwehr kommt: »Feuerwehr macht tatütataa. Feuerwehr hat Pumpen und Schläuche.« In anderen Bereichen kann es zunächst wenig zur Konversation beitragen oder Handlungen davon ableiten. Eltern wissen: Eine adäquate Reaktion auf »Räum dein Zimmer auf«, wird erst viel später erlernt.

Unterschied zwischen Lernen und Training

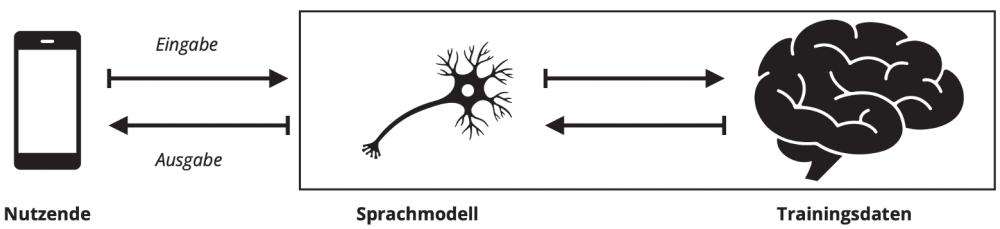

Abbildung 1.4: Funktion des Sprachmodells beim Abgleich von Trainingsdaten

Während Babys ihre Umgebung durch kontinuierliche Erfahrungen erkunden und *lernen*, indem sie auf Belohnungen und Enttäuschungen reagieren, funktioniert das Training von Sprachmodellen auf eine völlig andere Weise. Sprachmodelle werden ohne Pause mit riesigen Mengen an Textdaten *trainiert*.

✔ Dabei passt das Netzwerk seine Gewichte so an, dass es die statistischen Zusammenhänge in den Daten verallgemeinert und darauf aufbauend Texte generieren kann. Anstatt wie im Beispiel des Babys durch direkte Interaktion mit der Welt zu lernen, basiert das Wissen von Sprachmodellen allein auf den Texten, die sie verarbeitet haben.

- ✔ Es werden keine spezifischen Trainingsdaten gespeichert. Stattdessen repräsentieren die im Trainingsprozess Stück für Stück angepassten Parameter des Modells das gelernte Wissen (siehe *Abbildung 1.4*). Das Modell verarbeitet eine *Eingabe*, indem es sie in eine für Computer verständliche Form umwandelt.
- ✔ Diese Informationen werden dann durch verschiedene Ebenen des KNNs weitergeleitet, wobei jede Ebene die Daten ein Stück weiter verändert. Am Ende entsteht eine *Ausgabe*, die, gegeben der vormals gelernten statistischen Zusammenhänge, eine sinnvolle Antwort generiert (siehe *Abbildung 1.5*).

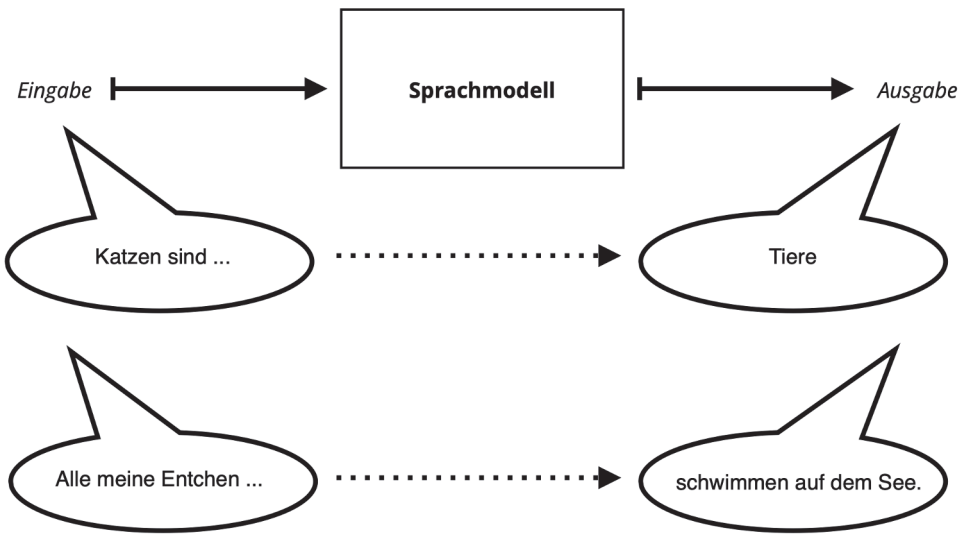

Abbildung 1.5: Verarbeitung von Nutzereingaben durch ein Sprachmodell

Woher stammen die Trainingsdaten?

OpenAI hat schon 2023 in einer Stellungnahme gegenüber dem britischen Parlament erklärt, es sei »unmöglich«, KI-Sprachmodelle wie ihr ChatGPT ohne Zugriff auf urheberrechtsgeschützte Materialien zu trainieren. Die Beschränkung auf urheberrechtsfreie Daten wäre »nicht mehr als ein interessantes Experiment«.

Selten wurden Urhebende gefragt

Werden wir in großem Stil beklaut und nicht an den Gewinnen der KI-Betreiber beteiligt? Diese Frage stellen sich Content Creators und sie ist Gegenstand zahlreicher Prozesse.

- ✔ Sollten Gerichte im Sinne der Rechteinhabenden entscheiden (siehe *Teil V*), kann die Löschung von KI-Modellen oder Produkten drohen oder es kommt zu starken Einschränkungen, und Lizenzkosten für Trainingsinhalte müssen bei Zeitungsarchiven, Buchverlagen oder Bilddatenbanken eingekauft werden.

- ✔ Bisher lief das noch völlig anders. Unter KI-Kennern gilt: Trau keiner KI, die du nicht selbst trainiert hast. Modelle werden mit gigantischen Daten ohne Herkunftsangabe gefüttert. In einigen Fällen heißt es schwammig, alle »frei verfügbaren« Daten seien eingeflossen. Das meint etwa Inhalte von YouTube oder Wikipedia.
- ✔ Sollten Sie mal eine Homepage aufgesetzt oder Postings mit Bildern oder Texten in Social Media veröffentlicht haben, ist die Wahrscheinlichkeit groß, dass auch mit Ihren Inhalten ohne Ihr Wissen und Einverständnis eine KI trainiert wurde. Wie Sie dem zumindest für die Zukunft widersprechen, erfahren Sie in *Kapitel 20*.

Nachträglich Quellen ergründen

Über Quellen und Inhalte von Trainingsdaten schweigen sich KI-Unternehmen aus, weil sie es entweder selbst nicht wissen oder den Rechtsstreit scheuen. Einige Datensätze kann man im Internet einsehen. Sie bestehen größtenteils aus dubiosen Internetquellen, fragwürdigen Inhalten mit ungeklärten Urheberrechten, Diskriminierung, ekelhafter (Kinder-)Pornografie und anderer Kriminalität. Da möchten einige Projekte Licht ins Dunkel bringen:

- ✔ Der Bayerische Rundfunk hat beispielhaft aufbereitet, was der deutsche Datensatz **LAION-5B** enthält und wieso es nahezu unmöglich ist, einmal gesammelte Daten entfernen zu lassen: `https://interaktiv.br.de/ki-trainingsdaten/`.
- ✔ **Spawning.ai** (`https://spawning.ai`) möchte informieren, wer Inhalte ohne Zustimmung verwendet, und bietet mit KI-Blockern, Browsererweiterungen und anderen Funktionen einen Ansatz dafür, diese Daten vom Training auszuschließen.

Mit **Have I Been Trained** (`https://haveibeentrained.com`) können Sie die gängigsten Datensätze durchsuchen, die zum Trainieren der KI-Systeme verwendet werden, um herauszufinden, ob Ihre Arbeit oder Fotos von Ihnen darin enthalten sind. Leider gibt es keine Möglichkeit, sich selbst aus einem bestehenden Datensatz zu entfernen, aber Sie können sich gegen künftiges Training entscheiden, indem Sie Ihre Bilder oder Domänen zur sogenannten »Do Not Train Registry« hinzufügen. Jedes Unternehmen, das diese Ausschlüsse (*Opt-outs*) berücksichtigt, wird dann die Links zu Ihren Bildern aus dem Datensatz entfernen, bevor es ein neues Modell trainiert.

GPT, was?

Die Abkürzung **GPT** steht für ***G**enerative **P**retrained **T**ransformer* und ist das derzeit bekannteste Sprachmodell. Die Bezeichnung erklärt auch, wie dieses Sprachmodell funktioniert, nämlich *generativ*, also Neues erschaffend, vortrainiert und mit *Transformer*-Architektur. Gehen wir die Buchstaben einmal im Einzelnen durch, um zu verstehen, was dort geschieht und was das Bahnbrechende an dieser Technologie ist.

G steht für generativ

Dass es einen Unterschied zwischen KI und generativer KI gibt, haben Sie bereits geahnt (vielleicht haben Sie »Künstliche Intelligenz für Dummies« bereits im Bücherregal). Im

Vergleich zu traditioneller KI ist ***generative KI*** nicht nur in der Lage, aus dem Datensatz den richtigen Wert herauszusuchen, sondern kann völlig neue Inhalte erzeugen. Es generiert also Neues. Eine Zwischenstufe wäre die *prädiktive,* also voraussagende, Fähigkeit, die jedoch ebenfalls auf vielen Berechnungen beruht. Auch wenn es für uns erst einmal unvorstellbar klingt, wie das möglich ist, müssen wir uns an diese Tatsache gewöhnen. Schwacher Trost: Die Entwickler dieser Technologie waren zunächst auch erstaunt, was da passiert.

Der Unterschied zwischen traditioneller und generativer KI

Eigenschaften	Traditionelle (prädiktive) KI	Generative KI
Hauptfunktion	Muster erkennen und Vorhersagen treffen (*prädiktiv*)	Originelle Inhalte *generieren*
Einsatzzweck	Analyse und Optimierung von Daten (*Big Data*)	Kunst, Text, Musik, Video, Programmierung
Kreativität	Kann keine neuen Inhalte erstellen	*Transformer-* (KNN mit Selbstaufmerksamkeit) oder *Diffusionsmodelle* erstellen Inhalte
Training	Spezifische Datensätze für die eingesetzten Aufgaben	Umfangreiche, vielfältige und unsortierte Datensätze
Anwendung	Sprach- und Bilderkennung, Analyse	Textgenerator, Bildgenerator, Codegenerator

P bedeutet pretrained (vortrainiert)

Ähnlich wie es Kinder beim Spielen machen, müssen KI-Modelle eine ganze Weile mit gewonnenen Informationen spielen, um sie zu verarbeiten. Und noch eine Situation könnte man anführen: Pianistinnen und andere Musizierende benötigen die berühmten »10.000 Stunden«-Übung, um ihre Finger ohne nachzudenken über ihr Instrument gleiten zu lassen.

Das KI-Training läuft in mehreren Stufen ab:

- ✔ Man unterscheidet das unüberwachte (*unsupervised*) Lernen und das überwachte (*supervised*) Lernen, das Zielvorgaben enthält.
- ✔ Was die generativen Systeme erstellen oder wiedergeben, haben sie nicht nur gelernt, sondern vielfach durchgespielt und immer wieder korrigiert (*backpropagation*).
- ✔ Training basiert auf schier unvorstellbar vielen Daten wie Literatur, Wikipedia oder aus Foreneinträgen – und da steht auch ziemlich viel Quatsch drin.
- ✔ Bei *vortrainierten Sprachmodellen* steckt außerdem sehr viel Anleitung und Nachjustierung drin. Dieses Training mit menschlicher Bewertung (*Reinforcement Learning from Human Feedback, RLHF*) ähnelt dem Clickertraining bei Hunden.

So wie Kinder den Unterschied zwischen Gut und Böse durch ihre eigenen Erfahrungen oder von Erziehungspersonen erlernen, ist die KI erst einmal unglaublich auskunftsfreundlich. Sie weiß, wie man Sprengstoff herstellt oder eine Bank überfällt und könnte Handlungsempfehlungen bereitstellen. Dass sie das nicht soll, lernt sie von Menschen, und speichert die moralische Bewertung.

T wie Transformer

Zum Schluss fehlt noch das T in GPT: *Transformer* steht nicht für die Spielzeugroboter, die sich in Fahrzeuge verwandeln können. Es ist eine von Google bereits 2017 vorgestellte Technologie, um nicht jede Kleinigkeit aus dem Training nachträglich bewerten zu müssen. Darin ist ein *Aufmerksamkeitsmechanismus* (*Selbstaufmerksamkeit* oder *Self-Attention*) enthalten, der die einzelnen Worte der Eingabe gewichtet und so **Sinnzusammenhänge aus dem Kontext** liest. Der macht das System nicht nur schneller, sondern auch effektiver im Erfassen von Zusammenhängen. ChatGPT und andere KI-Modelle basieren auf diesem ***Transformer-Modell*** beziehungsweise der Architektur.

Statt wie ältere KI-Systeme einen Text Wort für Wort durchzugehen, schaut der *Transformer* alle Wörter gleichzeitig an und bewertet, welche zusammengehören – so wie Sie beim Lesen eines Satzes oder einer Zeile sofort erkennen, welche Wörter sich aufeinander beziehen. Dafür berechnet der *Aufmerksamkeitsmechanismus* für jedes Wort einen Vergleichswert.

Im Training spricht man von *selbstüberwacht* (*self-supervised*). Erst das ermöglicht es, umfangreiche Daten ohne menschlichen Eingriff zu trainieren.

Richtig entrauschen für stabile Diffusion

Es gibt eine weitere Familie von Modellen, die besonders bei der Bild- und Videogenerierung (aber nicht ausschließlich) in Kombination mit der Transformer-Technologie zum Einsatz kommt: die Diffusion. Damit kann KI beispielsweise aus beschreibenden Worten ein Bild entwerfen (*Text-to-Image*). Vorteil: Diese Technik liefert hochmoderne Bildqualität. Prominente Beispiele sind Midjourney, Dall-E und Stable Diffusion, Letzteres hat das ja im Namen. Sie lernen alle in *Kapitel 10* kennen. Aber wie funktionieren diese Modelle?

Mathe und Magie: der Prozess der Diffusion

Diffusion kennen Sie aus dem Physikunterricht: Moleküle verteilen sich gleichmäßig im Raum und bringen ein System wieder ins homogene Gleichgewicht. Auf KI übertragen, wird erst ein diffuses Rauschen erzeugt (*Vorwärts-* oder *Diffusionsprogress*). Beim *Rückwärtsprozess* geschieht Magie: Scheinbar aus dem Nichts entstehen völlig neue Bilder und Inhalte. Eigentlich ist es Mathematik nach **Carl Friedrich Gauß**, der das Rauschen beschrieb. Sie kennen ihn vom Weichzeichner, dem *Gauß-Filter*, in der Bildbearbeitung.

Beispiel: Bleigießen

Stellen Sie es sich stark vereinfacht vor, wie Ihre Zukunftsvorhersage beim Bleigießen an Silvester. Sie kennen das: Sie erhitzen auf ihrem Löffel ein Stück Zinn (Blei ist das ja schon lange nicht mehr). Es schmilzt, und dann werfen Sie es schlagartig in kaltes Wasser. Das Zinn nimmt zufällige Formen an, und Sie können sie anhand von Erläuterungslisten oder Ihrer Vorstellung als Tierfiguren, Glücksbringer oder anderes identifizieren und deuten, ob Sie im nächsten Jahr die große Liebe oder doch nur den Lottogewinn finden.

So ähnlich macht es das Programm auch – mit dem Unterschied, dass es die Zwischenschritte vom geschmolzenen Zinn zum Aushärten im kalten Wasser so steuert, dass zum Beispiel am Ende ein schönes Alpaka herauskommt.

Denoising

Dieser magische Prozess ist auf den ersten Blick schwer nachzuvollziehen. Er funktioniert, indem das Modell im Training gezeigt bekommt, wie sich zuvor *verrauschte* Bilder – letztlich einfach ein Matsch von zufälligen Pixeln – in Zwischenschritten zurück zum Original wandeln.

- ✔ Man generiert eine gewisse Anzahl an Zwischenschritten vom verrauschten bis zum komplett *entrauschten* Bild. So als könnte man in einigen festgelegten Intervallen die Momente vom geschmolzenen Zinn zum Alpaka einfrieren und speichern.
- ✔ Macht man dies für sehr viele Beispiele im Training des Modells, kann das Modell mit den entsprechenden Lernalgorithmen daraus einstudieren, wie es sinnvolle Bilder aus einem verrauschten Bild durch *Entrauschen* konstruieren kann.

Der schrittweise Rückverwandlungsprozess bei Diffusion heißt *Denoising*.

Kombination mit dem Transformer

Wie weiß das Modell aber, welches Bild es ausgeben soll? Also ob es ein Alpaka oder doch lieber ein Lama generieren soll? Hier kommt die Kombination mit der Transformer-Technologie ins Spiel, die in diesem Fall nicht dafür verwendet wird, Wörter zu generieren, sondern die abstrakten Konzepte von Bildern und Texten zusammenbringt.

- ✔ Letztlich vergibt das Modell so etwas wie eine Punktzahl für ein gegebenes Bild-Text-Paar. Je besser Bild und Text zusammenpassen, umso höher die Punktzahl.
- ✔ Das Diffusionsmodell kann im nächsten Schritt den Vorgang so anpassen, dass das Zielobjekt im *Entrauschungsvorgang* näher an der gegebenen Beschreibung ist.
- ✔ Nach einer gewissen Anzahl an Wiederholungen entspricht das produzierte Bild einer Kombination aus Ihrer Eingabe und der gelernten Vorstellung des Modells.
- ✔ Mittels des sogenannten *Inpaintings* können innerhalb eines bereits generierten Bildes Elemente ausgetauscht oder verändert werden (in *Kapitel 10* lesen Sie darüber). Der *Entrauschungsvorgang* wird hier einfach mit den vorhandenen *maskierten* oder verrauschten Pixeln an dieser Stelle durchgeführt.

Bei Videofilmen kommt noch eine weitere Ebene hinzu, denn flüssig laufende Videobilder bestehen aus vielen Einzelbildern, die zusätzlich in einer Abfolge logisch zueinander passen müssen (*temporale Kohärenz*).

Die schnelle Einführung im YouTube-Video »**Generative AI in a Nutshell**« von Henrik Kniberg ist ein perfekter Start, wenn Sie dazu noch Fragen haben oder einen Gesamtüberblick brauchen (in englischer Sprache):
`https://www.youtube.com/watch?v=2IK3DFHRFfw`.