Dipl.-Ing. Hardo Naumann
Accellence Technologies GmbH
Garbsener Landstr. 10
D-30419 Hannover

Künstliche Intelligenz

KI - Grundlagen und Praxis

Status: in Arbeit

Inhalt

1 KI-Modelle

KI (engl. AI, Artificial Intelligence) steht als Abkürzung für **K**ünstliche **I**ntelligenz.

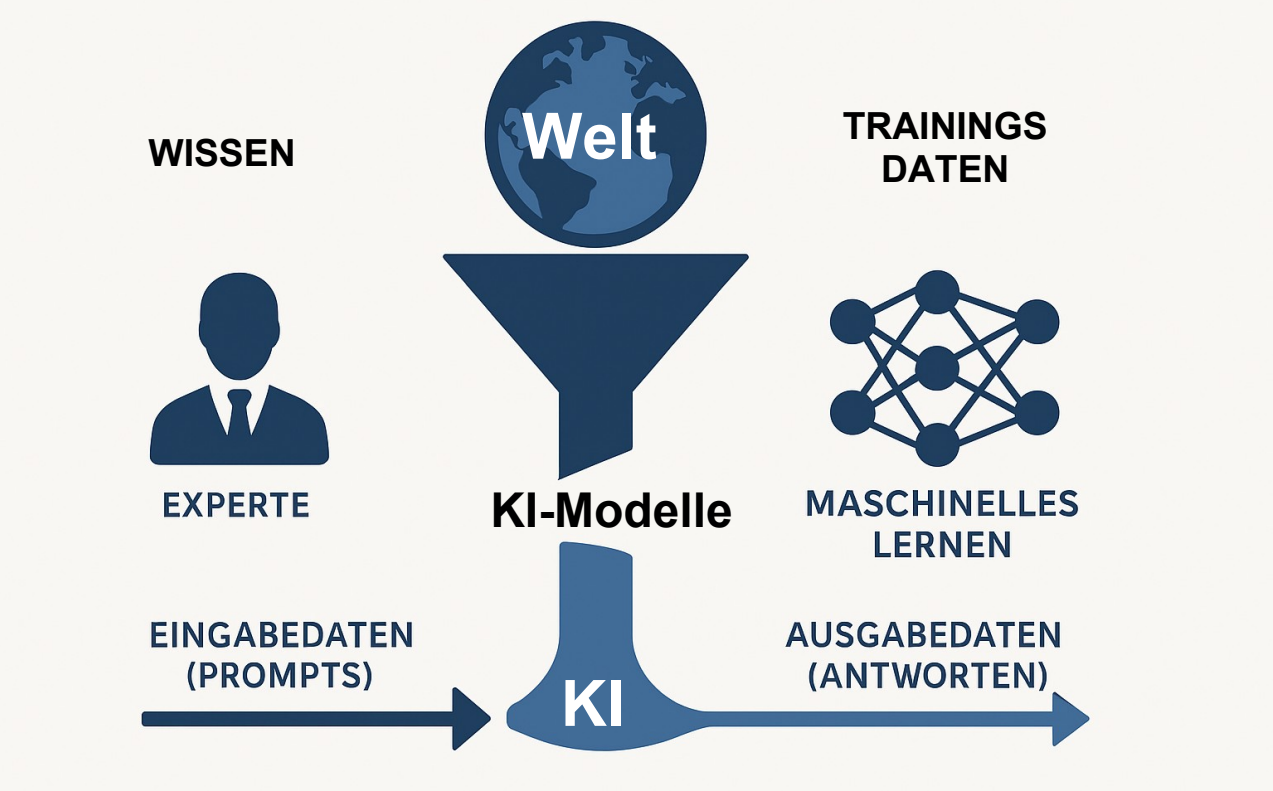
Als KI werden technische Systeme bezeichnet, die Aufgaben übernehmen können, die früher menschlicher Intelligenz vorbehalten waren.

Dazu arbeiten diese Systeme mit vereinfachten Abbildungen der Realität (**Modellen**). Diese Modelle werden in Form von Datenstrukturen und Rechenvorschriften implementiert, mit denen relevantes **Wissen** (z.B. Regeln, Zusammenhänge, Muster oder Wahrscheinlichkeiten) in einer für Maschinen verarbeitbaren stark verdichteten Form gespeichert wird. Auf dieser Basis können KI-Systeme Entscheidungen treffen, Vorhersagen ableiten oder Schlussfolgerungen ziehen.

Je nach Ansatz werden diese Modelle

- **wissensbasiert** durch Menschen erstellt (z.B. in **Expertensystemen**) oder
- **datengetrieben** durch **Maschinelles Lernen (ML)** weiterentwickelt

In der Praxis werden häufig beide Ansätze kombiniert (**Hybride KI**):



Ein KI-Modell besteht aus

- seiner Struktur (Form → **Architektur**) und
- dem darin gesammelten Wissen (Inhalt → **Daten**)

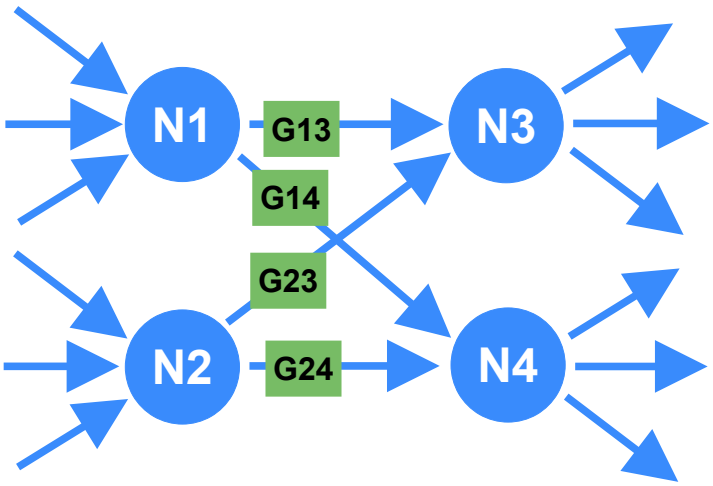
Die Architektur legt fest, wie Informationen verarbeitet werden können,
die Daten sind entscheidend dafür, was die KI „weiß“, d.h. über welche Informationen sie verfügt.

Expertenwissen wird in Datensätzen und Regeln **explizit** formuliert und ist für Menschen nachvollziehbar. ML-Systeme „lernen“ ihr Wissen dagegen durch Optimierung ihrer Modell-parameter aus Trainingsdaten **implizit**, was angesichts der großen Anzahl von Parametern (ggf. viele Milliarden) für Menschen schwer nachzuvollziehen ist.

2 Neuronale Netze

Beim Maschinellen Lernen (ML) werden künstliche Neuronale Netze verwendet.

Neuronale Netze bestehen aus Neuronen N mit gewichteten Verbindungen G:



Ein **Neuron** ist hier eine einfache Recheneinheit, die Zahlenwerte von anderen Neuronen empfängt, diese mit **Gewichten** multipliziert und aufsummiert. Das Ergebnis wird durch eine nichtlineare mathematische Funktion (**Aktivierungsfunktion**) geschickt und anschließend an andere Neuronen weitergegeben.

Nichtlineare Aktivierungsfunktionen sind erforderlich, damit Neuronale Netze auch komplexe Funktionen abbilden können. Außerdem müssen die Neuronen je nach Aufgabe in ausreichender Anzahl und Tiefe miteinander verbunden sein. Werden viele Ebenen solcher Neuronen hintereinander geschaltet, spricht man von **Deep Learning (DL)**.

- Ein Gewicht G > 0 verstärkt den Einfluss eines Signals auf die Weitergabe der Daten.
- Ein Gewicht G < 0 wirkt hemmend bzw. gegensätzlich.
- Ein Gewicht G = 0 bewirkt, dass der entsprechende Eingang über diese Verbindung keinen Beitrag zur Ausgabe des Neurons leistet.

Dieses Anpassen von Gewichten in Neuronalen Netzen wurde dadurch inspiriert, dass sich im menschlichen Gehirn Verbindungen durch Übung und Wiederholung festigen oder abschwächen; in analoger Weise können beim Deep Learning Verbindungen zwischen Neuronen mithilfe der Gewichte verstärkt oder abgeschwächt werden. Dies bildet aber nur einen stark vereinfachten funktionalen Aspekt biologischen Lernens ab.

Beim Training Neuronaler Netze werden die Gewichte systematisch so angepasst, dass Abweichungen zwischen der vom Netz erzeugten Ausgabe und der vorgegebenen **Zielfunktion** (z.B. einer Vorhersage) möglichst klein werden.

Für das Training wird nur ein Teil der verfügbaren Daten verwendet, denn weitere Daten werden anschließend zum Prüfen des Netzes benötigt, um sicher zu stellen, dass das Netz nicht nur mit den Trainingsdaten funktioniert, sondern auch mit Daten, die davon abweichen und dem Netz bisher unbekannt waren. Bleibt die Abweichung auch mit diesen Validierungsdaten zuverlässig ausreichend klein, dann ist das Netz austrainiert und kann verwendet werden.

Durch wiederholte Anpassung der Gewichte über viele Durchläufe (Iterationen) verbessert sich das Modell schrittweise und „lernt“ auf diese Weise statistische Regelmäßigkeiten und Muster aus den Trainingsdaten.

Die Gewichte stellen (im Gegensatz zum Expertenwissen) keine **expliziten** Regeln dar, sondern ergeben sich **implizit** aus statistischen Abhängigkeiten zwischen den im Training „gelernten“ Ein- und Ausgabedaten. Durch die große Anzahl von Parametern und Verzweigungen in viele Richtungen (**Dimensionen**), die in heutigen Neuronalen Netzen verarbeitet werden, kann KI ggf. auch Zusammenhänge berücksichtigen und nutzen, die Menschen in dieser Komplexität nicht auf Anhieb sehen oder verstehen.

Beim Maschinellen Lernen werden nicht die Trainingsdaten gespeichert, sondern nur die beim Training des Neuronalen Netzes ermittelten Gewichte. Diese Parameter geben statistische Zusammenhänge und Regelmäßigkeiten in den Trainingsdaten wieder und stellen somit eine stark komprimierte Abbildung (**Repräsentation**) der Verhältnisse der realen Welt dar.

Neuronale Netze aktueller KIs wie ChatGPT enthalten viele Milliarden solcher Gewichte, die in vielen Durchläufen angepasst werden müssen. Dieser massive Verdichtungsprozess der umfangreichen Daten aus der Realwelt erfordert deshalb viel Zeit und Rechenleistung. Dabei werden aus Tausenden von Beispielen die für die jeweilige Aufgabenstellung relevanten Regeln und Strukturen extrahiert und in einem vieldimensionalen Vektorraum (einem mathematischen Raum, in dem Inhalte nach vielen unterschiedlichen Kriterien geordnet werden können) so organisiert, dass ähnliche Daten nahe beieinander zu liegen kommen und Beziehungen zwischen den Daten als Vektoren (Pfeile von Punkt A nach B) ausgedrückt werden können. Diese Vektoren erlauben es, Informationen systematisch von einem Kontext in einen anderen zu transformieren, etwa von einer Sprache in eine andere. Auf diese Weise können bei Anfragen sehr effizient passende Ausgabedaten berechnet werden:

- Training Neuronaler Netze → großer Zeit- und Rechenaufwand
- Ausgabedaten für eine Anfrage ermitteln → schnell und effizient

KI kann auf diese Weise viele Aufgaben unterstützen und die Arbeit dadurch effizienter machen.

Durch die große Vielzahl von Anfragen kann aber auch der erforderliche Rechenaufwand für das Ermitteln der Ausgabedaten (Inferenzkosten) in Summe erheblich anwachsen und die Traininings-kosten bei weitem übersteigen.

Fazit:

- Jedes Neuron N ist ein einfaches Rechenelement, das die mit den Gewichten G multiplizierten Eingangsdaten summiert und über eine nichtlineare Funktion in Ausgangsdaten wandelt
- Das Neuronale Netz enthält nach dem Training die gefundenen Muster - aber nicht als lesbare Regeln, sondern verteilt in vielen einfachen Zahlen, nämlich allen ermittelten Gewichten G
- Die Gewichte stellen quasi die „Erfahrung“ des Neuronalen Netzes dar - in stark verdichteter Form als digitales Abbild (**Repräsentation**) des aus den Trainingsdaten „Gelernten“
- Einzelne Neuronen sind leicht verständlich - das Zusammenspiel von Tausenden oder Millionen solcher Neuronen bewirkt jedoch überraschende und erstaunliche Fähigkeiten → **Emergenz**

3 Intelligenz

Ist KI wirklich intelligent, oder simuliert sie nur Intelligenz?

Alan Turing schlug den nach ihm benannten Test [6] vor, um zu prüfen, ob eine Maschine über ein dem Menschen gleichwertiges Denkvermögen verfügt. Der Turing-Test misst jedoch ausschließlich beobachtbares Verhalten. Was er nicht berücksichtigt: Maschinen könnten menschliches Denken überzeugend simulieren – einschließlich aller Strategien, die erforderlich sind, um den Turing-Test zu bestehen – ohne dabei etwas wirklich zu „verstehen“.

Dies wird durch das sogenannte Chinese-Room-Argument [4] verdeutlicht, das zeigt, dass das bloße Befolgen festgelegter Regeln zum Verarbeiten von Zeichen und Zeichenfolgen nicht automatisch bedeutet, dass deren Bedeutung verstanden wird.

Je nach Aufgabenstellung können Antworten einer KI durchaus nützlich sein. Daraus kann jedoch nicht gefolgert werden, dass die KI das zugrunde liegende Problem wirklich „verstanden“, „reflektiert“ oder gar „bewusst“ bearbeitet hat.

Ein Gedankenexperiment soll dies verdeutlichen: Ein Roboter sei mit Sensoren und einer KI ausgestattet, die auf einen Sturz mit dem Ausruf „Aua“ reagiert. Daraus zu schließen, dass die KI Schmerz empfindet, wäre offenkundig naiv.

In gleicher Weise wäre es ein Trugschluss, aus funktional überzeugenden Antworten einer KI auf Verständnis oder gar Selbstbewusstsein zu schließen.

Trotzdem ziehen viele Menschen genau solche Schlüsse, begünstigt durch die sprachlich meist wohlklingenden Antworten moderner KI-Systeme, die dennoch beliebig falsch sein können.

Inzwischen mehren sich Berichte über emotionale Bindungen von Menschen an KI-Anwendungen [3]; die dabei erlebten Emotionen entstehen jedoch ausschließlich in den Köpfen der Nutzer.

Ein verbreiteter Denkfehler besteht darin, KI in Kategorien zu bewerten und Begriffe zu verwenden, die eigentlich Menschen vorbehalten sind. Die „Vermenschlichung“ (Anthropomorphisierung) von Maschinen trübt den Blick auf die tatsächlichen Eigenschaften und Funktionsweise; eine Personifizierung von KI verleitet zu falschen Erwartungen. Menschen neigen dazu, ihre eigenen Denk-, Gefühls- und Erklärungsmuster auf Maschinen zu projizieren. Zwar nahm die Entwicklung Künstlicher Intelligenz ihren Ausgangspunkt im Versuch, Strukturen menschlicher Intelligenz – etwa in Form Neuronaler Netze – nachzubilden; inzwischen hat sich die KI-Entwicklung jedoch in vielen Bereichen von ihrem biologischen Vorbild emanzipiert und nutzt Methoden, die sich deutlich vom menschlichen Denken unterscheiden.

Dies wurde erstmals im Kontext von AlphaGo deutlich sichtbar: Einige seiner Spielzüge erschienen menschlichen Kommentatoren zunächst sinnlos oder fehlerhaft, erwiesen sich jedoch im weiteren Verlauf als strategisch überlegen [5]. Während Menschen komplexe Aufgaben typischerweise in überschaubare Teilschritte zerlegen und sich dabei auf Heuristiken stützen, die oft auf individuellen Erfahrungen und (Vor-)Urteilen basieren, kann KI aufgrund hoher Rechenleistung und großer Speicherkapazitäten weitaus größere Lösungsräume systematisch durchforsten. Auf diese Weise findet sie Lösungen, auf die Menschen mit ihren Gewohnheiten und beschränkten Fähigkeiten nicht kommen.

Mit KI können viele Aufgaben sehr effizient gelöst werden, mitunter weitaus schneller und besser als Menschen dies könnten. Allerdings fehlt häufig ein tiefes Verständnis für die Realität, welches jeder menschlicher Entscheidung zugrunde liegt. Dies führt regelmäßig zu Fehlentscheidungen oder der Annahme von unrealistischen Kontexten.

Aktuelle Trends deuten darauf hin, dass diese Limitation grundsätzlich in den nächsten Jahren bestehen bleibt. Eine umfassende Lösung dieser Schwachstellen würde bedeuten, dass KI auf dem gleichen oder höherem Niveau von Menschen handeln und dessen Aufgaben nahezu vollständig übernehmen könnte.

Deshalb ist ein differenzierter Blick auf KI geboten: Bei manchen Aufgaben ist sie der menschlichen Leistungsfähigkeit weit überlegen, in anderen ist sie vollkommen ungeeignet, und häufig liegt sie zwischen diesen beiden Polen.

Aus Sicht der Praxis ist es nicht wichtig, ob KI „intelligent“ ist und Zusammenhänge „versteht“, sondern ob sie die gestellten Aufgaben zuverlässig erledigt. Ein „Zuviel“ an Intelligenz bedeutet da eher ein Risiko.

Die Herausforderung besteht deshalb darin, Anwendungsfelder und Nutzungsmodelle zu finden, in denen KI sinnvoll eingesetzt werden kann, und zugleich die mit ihrem Einsatz verbundenen Risiken zuverlässig einzuhegen.

4 Multidimensionalität

Woher kommt die (scheinbare) Intelligenz von KI-Systemen?

Um diese Frage zu beantworten, lohnt sich ein Blick auf die **Entwicklung der Wissenschaft**.

Forschung ist im Wesentlichen die Suche nach den Regeln, nach denen unsere Welt funktioniert. Wenn man Zusammenhänge versteht, Regeln erkennen und idealerweise als mathematische Formeln formulieren kann, können fehlende oder künftige Werte zuverlässig berechnet werden. Wir können heute auf dieser Basis auf Jahre im Voraus mit recht hoher Genauigkeit berechnen, mit wie vielen Kilometern Abstand ein Asteroid an der Erde vorbeifliegen wird.

Entscheidend für viele wissenschaftliche Durchbrüche war ein **Perspektivwechsel**:

1. Die Flugbahn von Körpern im Schwerefeld der Erde wird erst dann als **Parabel** berechenbar, wenn Einflüsse wie Reibung und Wind gezielt ausgeblendet werden
2. Von der Erde aus erscheinen **Planetenbahnen** kompliziert und verworren, erst mit der Sonne als Bezugspunkt werden sie wesentlich einfacher und verständlicher
3. Erst wenn die Geschwindigkeit des **Lichts** als konstant angenommen und stattdessen Raum und Zeit als veränderlich betrachtet werden, können Bewegungen bei hohen Geschwindigkeiten präzise beschrieben werden

Diese Beispiele zeigen: Ob in Daten Strukturen und Zusammenhänge erkannt werden können, hängt oft entscheidend von der **Perspektive** ab, aus der sie betrachtet werden.

Genau darauf beruht auch ein wesentlicher Teil moderner KI: Durch die Transformation von Daten in höherdimensionale Merkmalsräume werden Zusammenhänge sichtbar, die in der ursprünglichen Darstellung verborgen bleiben.

Um dieses Prinzip anschaulich zu verdeutlichen, habe ich ein interaktives 3D-Modell entwickelt:

→ www.hardo-naumann.de/3D

Darin sind Daten als **farbige Punkte** in einem 3-dimensionalen Raum dargestellt.
Zur Orientierung dienen die **Koordinatenachsen x, y und z**.

Mit den Schiebereglern können Sie die Blickrichtung frei verändern und das Modell aus unterschiedlichen Perspektiven betrachten.

- **Können Sie regelmäßige Muster in diesem Modell entdecken?**

Die Buttons unter den Schiebereglern fahren automatisch die jeweils passende Perspektive an. Die damit sichtbaren Muster stehen symbolisch für die drei zuvor beschriebenen wissenschaftlichen Erkenntnisse.

Wie in der Wissenschaft werden auch bei KI-Systemen viele Zusammenhänge erst dann sichtbar, wenn die Daten aus dem richtigen Blickwinkel betrachtet werden.

Neuronale Netze lernen letztlich nichts anderes, als Daten schrittweise in neue Darstellungen zu überführen, in denen die gesuchten Muster leichter erkennbar werden. Die scheinbare Intelligenz besteht also hauptsächlich darin, eine geeignete Perspektive auf die Daten zu finden.

Neuronale Netze verfolgen also einen ähnlichen Ansatz wie die traditionelle Forschung:

- Sie sortieren Daten in vielen Dimensionen und
  suchen aus vielen unterschiedlichen Perspektiven nach Mustern darin
- Weil sie dabei jedoch Tausende Dimensionen berücksichtigen,
  finden sie auch Strukturen und Zusammenhänge,
  die unserem Vorstellungsvermögen bislang verborgen sind

Weil es so wichtig ist: Ein kleiner Exkurs zu den Dimensionen

Damit diese interaktive Analogie anschaulich und leicht nachvollziehbar bleibt, habe ich

- **nur 3 Dimensionen** verwendet.

Jeder Punkt in diesem Raum lässt sich durch einen Vektor mit den 3 Koordinaten (x, y, z) beschreiben.

Moderne Neuronale Netze arbeiten dagegen häufig mit

- **mehreren tausend Dimensionen**

Während des Trainings entstehen dabei hochdimensionale Repräsentationen der Daten, in denen

- zusammengehörige Datenpunkte nahe beieinander liegen,
- ähnliche Inhalte ähnliche Positionen einnehmen,
- Beziehungen durch Vektoren beschrieben werden können.

Solche Vektoren ermöglichen Transformationen zwischen unterschiedlichen Kontexten, beispielsweise die Übersetzung eines Textes von einer Sprache in eine andere. Für das Neuronale Netz erscheinen dabei sowohl natürliche Sprache als auch Softwarecode letztlich als Folgen von Vektoren in einem hochdimensionalen Raum.

Vergleich Mensch - Maschine:

- 2 Dimensionen können wir leicht erfassen, etwa eine Skizze auf einem Blatt Papier.
- 3 Dimensionen erfordern bereits ein gutes räumliches Vorstellungsvermögen.
- 4-dimensionale Raumzeit der Relativitätstheorie ist deutlich schwieriger zu begreifen.
- Große Neuronale Netze arbeiten mit mehreren tausend Dimensionen und entziehen sich daher weitgehend unserer direkten Vorstellungskraft.

Das interaktive 3D-Modell veranschaulicht das **„Geheimnis“ moderner KI**:

- Was auf den ersten Blick chaotisch aussieht,
  kann - aus der richtigen **Perspektive** betrachtet -
  einfachen Regeln folgen
- Neuronale Netze können - dank hoher Rechenleistung -
  Daten in Tausenden **Dimensionen** analysieren
  und dadurch **Strukturen und Muster** finden,
  die Menschen nicht so leicht sehen

→ KI ist nicht hochintelligent, sondern hochdimensional

Aus der Tatsache, dass KI-Systeme mit einer zwar sehr großen, aber dennoch endlichen Anzahl von Parametern eine ungleich größere Vielfalt an Ergebnissen erzeugen können, ziehe ich den Schluss: Viele Dinge der Welt können auf einfachere (und möglicherweise domänenübergreifend gemeinsame) Regeln und Strukturen zurückgeführt werden.

Ich vermute, dass Menschen diese Strukturen bislang deshalb nicht erkennen konnten, weil unser Vorstellungsvermögen auf wenige Dimensionen beschränkt ist, während Neuronale Netze dieselben Daten in Räumen mit Tausenden von Dimensionen analysieren können.

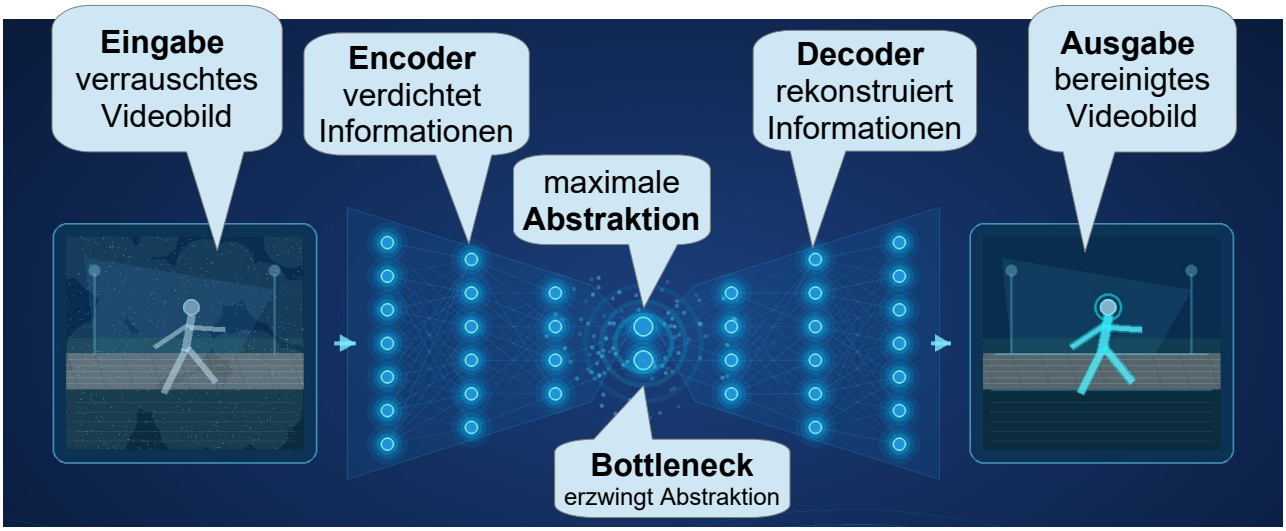
Möglicherweise liegt eine der größten Chancen der KI daher nicht nur in der Automatisierung von Aufgaben, sondern auch darin, bislang verborgene Gesetzmäßigkeiten sichtbar zu machen. Langfristig könnte es gelingen, aus den in Neuronalen Netzen gespeicherten Gewichten und Strukturen neue Erkenntnisse über die zugrunde liegenden Regeln unserer Welt abzuleiten.

5 Abstraktion

Menschliches Denken basiert auf verschiedenen Abstraktionsebenen. Beim Verstehen von Sprache abstrahieren wir beispielsweise von einzelnen Buchstaben über Wortschatz, Grammatik, Bedeutung und Stil bis hin zu übergeordneten Konzepten. Wir denken normalerweise nicht in einzelnen Buchstaben, sondern in immer stärker verdichteten Bedeutungseinheiten.

Neuronale Netze arbeiten ähnlich: Sie transformieren konkrete Eingabedaten schrittweise in zunehmend abstrakte Repräsentationen. Im Unterschied zum Menschen entsprechen diese Repräsentationen jedoch nicht zwangsläufig bekannten Kategorien wie Syntax oder Semantik, sondern entstehen während des Trainings selbstständig.

Besonders deutlich wird dies bei sogenannten **Autoencodern**:



Autoencoder sind leicht zu trainieren, denn die **Zielfunktion** besteht lediglich darin, dass die Abweichung der Ausgabedaten von den Eingabedaten minimiert werden soll, während das Neuronale Netz mit seiner **Architektur** erzwingt, dass alle Daten durch eine sehr schmale Zwischenschicht (Engstelle, Bottleneck) mit nur wenigen Neuronen Breite hindurch müssen.

Auf diese Weise wird das Neuronale Netz darauf trainiert, die Eingabedaten schrittweise maximal zu verdichten, bis im Bottleneck nur noch die wesentlichen Merkmale enthalten sind. Diese kompakte Repräsentation kann als eine Art hochabstrakte „Quintessenz“ der ursprünglichen Daten verstanden werden, aus der sich die Ausgangsdaten anschließend wieder näherungsweise rekonstruieren lassen. Als „Nebeneffekt“ werden bei einem optimal angepassten Autoencoder unerwünschte Details (z.B. Rauschen) aus den Daten entfernt, während die wesentlichen Inhalte klarer sichtbar werden.

Vermutlich beruht ein wesentlicher Teil der Leistungsfähigkeit moderner KI genau auf dieser Fähigkeit, komplexe Sachverhalte in sehr kompakte und zugleich informationsreiche Repräsentationen zu überführen. Auf diesen höheren Abstraktionsebenen werden konkrete Einzelheiten zunehmend ausgeblendet, während gemeinsame Strukturen und Zusammenhänge hervortreten. Dies könnte der Schlüssel dazu sein, mit einer vergleichsweise kleinen Anzahl allgemeiner Prinzipien eine große Vielfalt konkreter Aufgaben beschreiben und lösen zu können.

Das Prinzip der Abstraktion ist ähnlich wie beim menschlichen Denken; die Kategorien und Ebenen, mit denen KI Daten abstrahiert, sind jedoch ganz andere und bilden sich erst beim Training des Neuronalen Netzes heraus, indem sich ihre Eignung gegenüber anderen Schemata von Kategorien und Ebenen durch minimale Abweichungen der Ausgabedaten erweisen muss.

Die eigentliche Leistung der KI liegt möglicherweise nicht im „Denken“, sondern in der Fähigkeit, Informationen auf andere, für Menschen ungewohnte Weise zu abstrahieren und zu organisieren.

6 Tensoren

6.1 Repräsentation komplexer Datenstrukturen

Zum präziseren Verständnis Neuronaler Netze ist es wichtig zu wissen, wie unterschiedliche Datenarten in diesen Netzen verarbeitet werden.

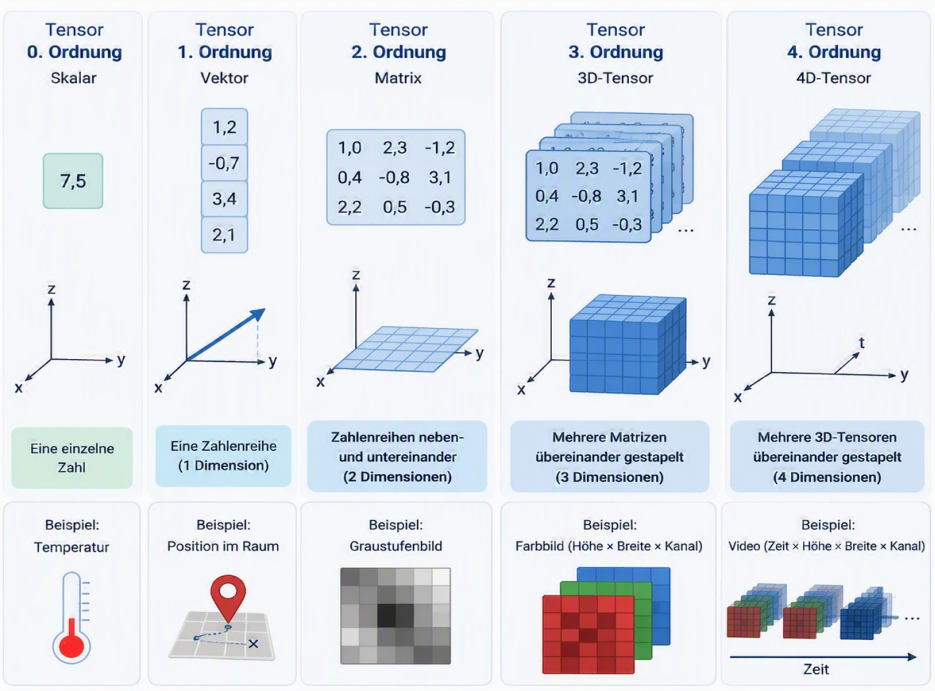
Ein Neuronales Netz, das ausschließlich einzelne Zahlen verarbeitet, wäre leicht zu verstehen und zu implementieren, aber nur sehr eingeschränkt nutzbar.

Um beliebige Datenarten verarbeiten zu können, führen wir den Begriff des Tensors ein. Tensoren dienen der universellen Darstellung (Repräsentation) von Daten innerhalb Neuronaler Netze.

Hinter diesem zunächst geheimnisvoll klingenden Begriff verbirgt sich ein einfaches Konzept: Ein Tensor fasst mehrere Zahlen zu einer gemeinsamen Struktur zusammen.

- Tensor 0. Ordnung: eine einzelne Zahl (Skalar)
- Tensor 1. Ordnung: eine Zahlenreihe (Vektor)
- Tensor 2. Ordnung: Zahlenreihen neben- und untereinander (Matrix)
- Tensor 3. Ordnung: mehrere Matrizen übereinander gestapelt
- Tensor 4. Ordnung: mehrere Tensoren 3. Ordnung
- usw.

Mit Tensoren lassen sich praktisch beliebige Datenarten darstellen:



Ein Graustufenbild kann beispielsweise als Tensor 2. Ordnung gespeichert werden, bei dem jeder Zahlenwert die Helligkeit eines Bildpunktes beschreibt.

Ein Farbbild enthält zusätzlich die Farbkanäle Rot, Grün und Blau und wird deshalb häufig als Tensor 3. Ordnung dargestellt.

Ein Video besteht aus einer Folge von Bildern und kann als Tensor 4. Ordnung aufgefasst werden, bei dem zusätzlich die Zeit als weitere Dimension hinzukommt.

Auch Sprache, Audiodaten oder Messwerte lassen sich auf diese Weise in Tensoren überführen und dadurch von Neuronalen Netzen verarbeiten.

Aus Sicht eines Neuronalen Netzes werden die „Rohdaten“ in Tensoren bereitgestellt. Durch die Verarbeitung in den verschiedenen Netzschichten werden diese Tensoren schrittweise in immer abstraktere Tensoren überführt. Ähnlich wie Menschen konkrete Wahrnehmungen zu Begriffen, Konzepten und Zusammenhängen abstrahieren, bilden Neuronale Netze zunehmend kompaktere und informationsreichere Repräsentationen der Daten. Der Unterschied besteht darin, dass diese Abstraktionsebenen nicht von Menschen definiert werden, sondern während des Trainings selbstständig entstehen.

6.2 Interaktives Lernen mit TensorFlow

Viele Eigenschaften Neuronaler Netze lassen sich leichter verstehen, wenn man ihnen beim Lernen direkt zusieht. Eine gute Möglichkeit dazu bietet der interaktive **TensorFlow Playground**. Dort können Sie ohne Programmierkenntnisse die Anzahl der Neuronen und Schichten verändern, verschiedene Datensätze auswählen und beobachten, wie sich die Gewichte und Entscheidungsgrenzen des Netzes während des Trainings verändern. Die Dicke der Verbindungslinien zeigt die Gewichte der einzelnen Verknüpfungen zwischen den Neuronen. Besonders anschaulich wird dabei, wie zusätzliche Schichten und Neuronen die Fähigkeit des Netzes verbessern, komplexe Muster zu erkennen und Daten schrittweise auf immer höhere Abstraktionsebenen zu transformieren.

Wer die Funktionsweise Neuronaler Netze wirklich verstehen möchte, sollte sich etwas Zeit nehmen und selbst damit experimentieren. Der Lerneffekt ist oft größer als nach vielen Seiten theoretischer Beschreibung. Der TensorFlow Playground ist kostenlos im Browser verfügbar:

→ https://playground.tensorflow.org

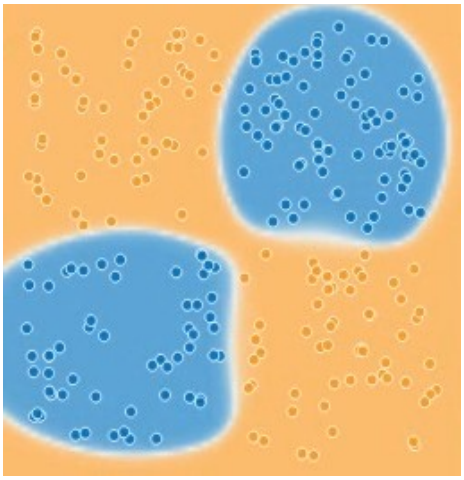
Wer nicht so viel Zeit aufwenden möchte, kann sich eine kurze Videosequenz anschauen, die ich mit TensorFlow erstellt habe:

→ https://www.hardo-naumann.de/video/tensorflow.mp4

Die **Trainingsdaten** bestehen aus **blauen** und **orangen** Punkten.

Gesucht ist eine Funktion, die den Hintergrund so einfärbt, dass die Farbe der Punkte anhand ihrer Position vorhergesagt werden kann. Menschen erkennen bei diesem Beispiel oft schon auf auf den ersten Blick, dass eine Aufteilung in 4 Quadranten mit jeweils wechselnden Farben eine gute Lösung wäre.

Das Neuronale Netz erreicht nach einigem Ausprobieren einen Fehler von 0 %, also keine Abweichung von der Zielfunktion. Das heißt, dass es eine Funktion gefunden hat, die für alle bekannten Trainingspunkte das vorgegebene Ergebnis liefert.



Dennoch bleiben am Rand einige Bereiche mit unsicheren oder falschen Vorhersagen. Anschaulich übertragen auf ein LLM entsprechen solche Bereiche Situationen, in denen das Modell für bestimmte Eingaben keine ausreichend passende Repräsentation aus den Trainingsdaten ableiten kann. Das entspricht genau den Stellen, bei denen ein LLM **fehlerhafte Antworten** oder **Halluzinationen** liefern würde.

Mit weiteren Trainingsdaten können diese Bereiche häufig verkleinert werden. Zu viele Trainingsdurchläufe oder ungeeignete Trainingsdaten können jedoch auch zu Overtraining (**Overfitting**) führen. Bei Overfitting lernt das Neuronale Netz die Position und Farbe aller Trainingspunkte nahezu auswendig. Die bekannten Punkte werden dann zwar weiterhin korrekt klassifiziert, die dazwischenliegenden Flächen werden jedoch nicht mehr als zusammenhängende Bereiche erkannt. Das Modell verliert dadurch die Fähigkeit, von den bekannten Punkten auf ähnliche, aber bisher unbekannte Daten zu verallgemeinern.

- *Gute KI lernt Strukturen und Beziehungen und kann dadurch generalisieren*
- *Overfitting bedeutet, dass statt der Struktur nur die Beispiele gelernt werden*

  Analogie zum Menschen: Auswendiglernen statt Zusammenhänge verstehen

6.3 Abstraktion mit Tensoren

Das Konzept der Abstraktion (Kapiel 5) lässt sich an den Tensoren sehr gut beobachten:

Die konkreten Ein- und Ausgabedaten haben in der Regel nur eine kleine Anzahl Dimensionen, aber viele Einzelwerte. Ein Full-HD-Videobild hat z.B. nur 3 Dimensionen (x, y, color), enthält aber Werte von ca. 2 Millionen Bildpunkten (Pixeln).

Im Zuge der Abstraktion steigt in der Regel die Anzahl der Dimensionen.

Bilder werden dann nicht mehr in den Dimensionen

- x-Koordinate
- y-Koordinate
- Farbe

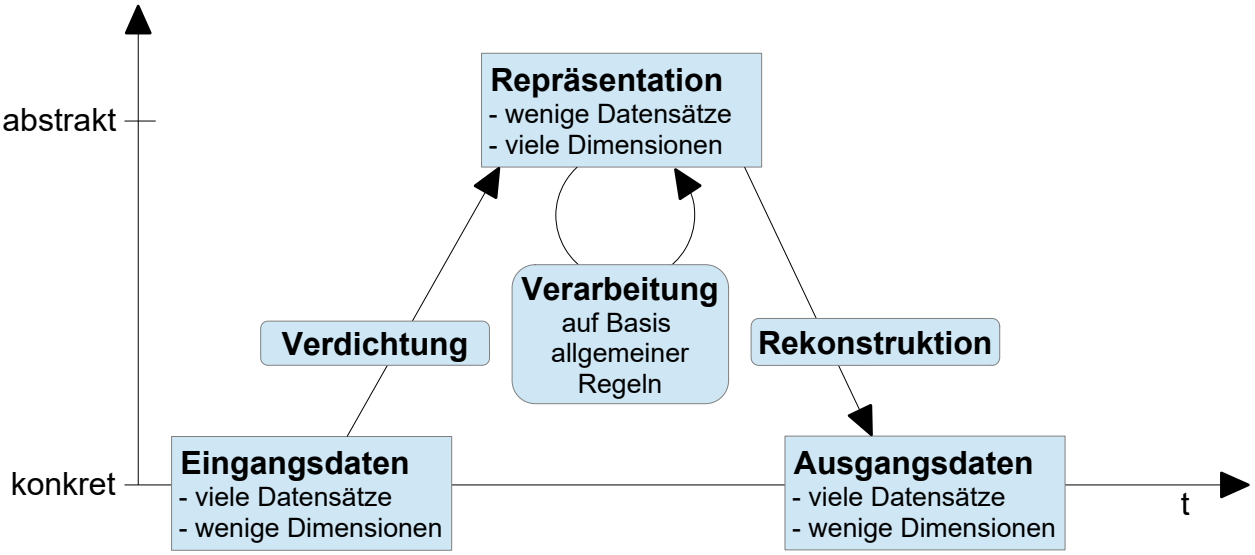
dargestellt, sondern mit abstrakten Merkmalen wie:

- Kanten
- Strukturen
- Bewegungen
- Texturen
- Objekte
- Beziehungen zwischen Objekten

und vielen weiteren, für Menschen oft nicht interpretierbaren Eigenschaften.

Mit steigender Anzahl Dimensionen reduziert sich im Gegenzug in der Regel die Anzahl der Datensätze: Statt durch 2.073.600 Pixel wird das Bild dann z.B. nur noch durch 100 Objekte repräsentiert, die dafür mehr Dimensionen aufweisen.

Die Rohdaten der realen Welt bestehen häufig aus einer sehr großen Anzahl konkreter Messwerte, beispielsweise Pixeln eines Bildes oder Abtastwerten eines Audiosignals. Neuronale Netze transformieren diese Daten schrittweise in zunehmend abstrakte Repräsentationen. Das entspricht dem in Kapitel 4 beschriebenen Perspektivwechsel: Aus wenigen ursprünglichen Dimensionen werden **hochdimensionale Merkmalsräume**, in denen nicht mehr einzelne Messwerte, sondern ihre Beziehungen und Zusammenhänge repräsentiert und besser erkennbar werden. Komplexe Sachverhalte können hier oft durch überraschend kompakte abstrakte Repräsentationen beschrieben und entsprechend effizient verarbeitet werden:



6.4 Daten mit Beziehungen

Warum verwenden wir Tensoren? Wir könnten sämtliche Farbwerte eines Videobildes auch in einer einfachen Zahlenreihe hintereinanderschreiben. Beim Speichern und Übertragen von Videobildern wird das auch genau so gemacht: Eine simple Zahlenreihe mit sämtlichen Pixelwerten, etwa als RGB oder YUV, würde ein Bild vollständig repräsentieren. Das Bild könnte daraus wieder exakt hergestellt werden (sofern bekannt ist, wieviele Pixel zu einer Zeile gehören; diese Information wird typischerweise als Metadaten im Header der Bilddaten gespeichert).

Bei so einer rein „linearen“ Organisation der Daten in nur 1 Dimension wären aber die räumlichen Beziehungen zwischen den Pixeln nicht mehr unmittelbar verfügbar – diese Beziehungen müssten unter Berücksichtigung der Metadaten erst aufwändig wieder rekonstruiert werden.

Pixel, die in 2 Zeilen direkt übereinander liegen und somit geometrisch eng zusammengehören, würden bei einem Full-HD-Videobild in der linearen Speicherorganisation 1920 Speicherstellen voneinander entfernt liegen – ihr Zusammenhang ginge verloren. Erst die **Repräsentation als Tensor** ermöglicht es, diese Pixel auch in ihrem digitalen Abbild unmittelbar „nebeneinander“ anzuordnen und ihre räumliche Beziehung für die weitere Auswertung optimiert bereitzustellen.

Dasselbe trifft auf die zeitliche Folge von Pixeln zu: Bei 1-dimensionaler linearer Repräsentation, wie beispielsweise im RAM oder bei einer seriellen Fernsehübertragung mit SDI, ist derselbe Pixel eines Full-HD-Bildes 2.073.600 Pixel-Speicherstellen von seinem zeitlich nur 40 Millisekunden früheren Zustand entfernt. Die wichtige Information, dass es derselbe Pixel ist, nur sehr kurze Zeit später, ist nicht mehr unmittelbar verfügbar.

Diese Beispiele zeigen, dass die Bedeutung von Tensoren für KI weit über die Erklärung

Ein Tensor ist eine Anordnung von Zahlen entlang n Achsen [8]

hinausgeht: Diese Definition ist mathematisch korrekt, erklärt aber nicht, warum Tensoren für KI so wichtig sind. Deshalb ergänze ich

Tensoren enthalten nicht nur Zahlen, sondern auch Beziehungen zwischen den Werten

Bei der Repräsentation als Tensor werden die Beziehungen zwischen den Daten unmittelbar in der Struktur der Daten abgebildet. Sie müssen deshalb nicht erst aus einer linearen Folge von Werten rekonstruiert werden. Tensoren spielen eine zentrale Rolle bei der "Organisation" der Daten, um innere Beziehungen zwischen den Daten effizient abbilden zu können.

Bei Videosequenzen sind Beziehungen zusammengehöriger Bildpunkte in mehreren Dimensionen intuitiv: sie können 1) nebeneinander liegen, 2) übereinander liegen, 3) zeitlich aufeinander folgen. Aber auch bei anderen Daten ist es von Vorteil, ihre Beziehungen in ggf. sehr vielen Dimensionen speichern und damit rechnen zu können.

Um das anschaulich zu vermitteln, nehmen wir als fiktives Denkmodell an, dass verschiedene Dimensionen unterschiedliche Arten von Beziehungen repräsentieren. So könnte es beispielsweise nützlich sein, Token von Sprachen in Tensoren so anzuordnen, das verschiedene Sprachen über eine der Dimensionen ausgewählt werden können, in einer anderen Dimension liegen vielleicht Synonyme direkt nebeneinander, während wieder eine andere Dimension auf einer Skala von Minus zu Plus Gegensätze und Steigerungsstufen desselben Ausdrucks bereitstellt. In realen Neuronalen Netzen sind diese Beziehungen jedoch meist auf viele Dimensionen gleichzeitig verteilt und nicht unmittelbar einzelnen Achsen zuordenbar.

Tensoren ermöglichen es, diese Beziehungen auch in sehr vielen Dimensionen gleichzeitig darstellen und auswerten zu können. Damit können auch komplexe Zusammenhänge erkannt und genutzt werden, die sich der menschlichen Intuition nicht unmittelbar erschließen.

Tensoren ermöglichen also nicht nur die Speicherung komplexer Datenstrukturen. Sie enthalten zugleich wichtige räumliche, zeitliche und semantische Beziehungen zwischen den Daten. Dadurch bilden sie die Grundlage dafür, dass Neuronale Netze Muster, Objekte, Bewegungen und Zusammenhänge überhaupt erkennen können.

Jede zusätzliche Tensor-Dimension ermöglicht die Darstellung einer weiteren Art von Beziehung zwischen den Daten. Das Prinzip „Abstraktion“ bedeutet auf dieser Basis: weniger konkrete Einzelheiten, mehr Beziehungen und Strukturen. Dadurch werden Zusammenhänge sichtbar, die in den Rohdaten nur schwer zu erkennen sind.

6.5 Unterschied von KI zur „klassischen“ Informatik

Informatiker könnten an dieser Stelle einwenden, dass eine Repräsentation der Beziehungen und Dimensionen mit eindeutig deklarierten **Objektstrukturen**, beispielsweise mit XML oder JSON, besser wäre als ein Tensor mit seinen **unbenannten (anonymen) Dimensionen**.

Allerdings erlauben genau diese anonymen Dimensionen Neuronalen Netzen überhaupt erst die erforderliche Freiheit zur Bildung optimierter Repräsentationen. Während des Trainings können dadurch dynamisch neue Kategorien und Abstraktionsebenen entstehen, die nicht von Menschen vorgegeben wurden, sondern sich aus den Trainingsdaten und der Zielfunktion ergeben.

Für Menschen sind Objektstrukturen mit benannten Kategorien besser geeignet:

- leicht verständlich
- leicht dokumentierbar
- leicht zu überprüfen

Der Nachteil ist:

- Entwickler müssen vorab wissen, welche Kategorien relevant sind
- Die Struktur wird von außen vorgegeben

Damit wird bereits festgelegt, welche Kategorien überhaupt berücksichtigt werden können.
Das würde das Modell auf die vom Entwickler vorgegebene Sicht der Daten beschränken.

Der besondere Nutzen Neuronaler Netze liegt darin, dass sie die passenden Kategorien mit Hilfe der Zielfunktion aus den Daten selbst ableiten:

Klassisches Datenmodell	Neuronale Netze
Mensch	Trainingsdaten
↓	↓
definiert Kategorien	Repräsentation auf Zielfunktion optimieren
↓	↓
ordnet Daten ein	dazu passende Kategorien bilden

Damit können Repräsentationen entstehen, die Menschen so nicht formuliert hätten, die sich für die jeweilige Aufgabe jedoch als besonders geeignet erweisen.

> *Klassische Informatik beschreibt die Welt in vorgegebenen Kategorien.*
> *Neuronale Netze leiten ihre Kategorien aus den Daten selbst ab.*

Die Kehrseite der Medaille ist allerdings, dass Menschen diese **optimierten Repräsentationen** häufig nicht mehr unmittelbar verstehen können. Das liegt nicht daran, dass KI eine „Black Box“ wäre – wir können jeden Zahlenwert, jede Gewichtung und jede Rechenoperation eines Neuronalen Netzes prinzipiell auslesen und nachvollziehen. Die eigentliche Herausforderung besteht vielmehr in der schieren Komplexität der dabei entstehenden Strukturen.

Vielleicht gelingt es uns eines Tages – möglicherweise sogar mit Hilfe einer KI –, die in diesen Tensoren und Repräsentationsräumen codierten Beziehungen und Regeln in eine für Menschen verständliche Form zu übersetzen oder zumindest überblicksweise zu visualisieren.

7 Historische Einordnung

Manches wird vielleicht klarer, wenn man die großen Entwicklungslinien der letzten Jahre verfolgt:

1. Die Entwicklung von **Suchmaschinen** über Semantic Web zu LLMs
2. Die Entwicklung der **Videokompression** von JPEG, MPEG, HEVC, Autoencoder, ...

Diese Entwicklungslinien können helfen, den aktuellen Stand von KI besser verstehen und einordnen zu können. Dieses Kapitel könnte auch heißen: **Von der Suchmaschine zum LLM**

Auf diese Weise wollen wir nun die zentrale Frage angehen:

- **Wie funktioniert KI eigentlich?**
- Konkret: Wie kann es sein, dass ein Large Language Model in Sekundenbruchteilen **umfassende Antworten auf nahezu beliebige Fragen** liefern kann?

Schauen wir dazu zuerst auf die Entwicklung der Suchmaschinen.

Der erste Ansatz war, einen Suchbegriff einzugeben und anschließend alle bekannten Webseiten nach diesem Begriff zu durchsuchen. Das dauert sehr lange, weil nach Eingabe der Suchanfrage erst alle Daten aus den Websites geladen und durchsucht werden müssen.

Der naheliegende nächste Entwicklungsschritt war deshalb, bereits vor der ersten Suchanfrage einen **Index** aufzubauen. Das Internet wird dazu regelmäßig durchsucht („gecrawlt“), und die gefundenen Informationen werden in einer geeigneten Datenstruktur organisiert. Diese Datenstruktur wird kontinuierlich optimiert, um zu jedem Suchbegriff möglichst schnell alle relevanten Informationen finden zu können.

Anfangs handelt es sich dabei um einfache Indizes, bei denen Suchbegriffe exakt übereinstimmen müssen. Später entwickelt sich daraus das **Semantic Web** beziehungsweise die semantische Suche. Nun müssen Wörter nicht mehr identisch sein – es genügt, wenn sie auf einer abstrakteren Ebene dieselbe Bedeutung oder dasselbe Themengebiet beschreiben.

In der nächsten Stufe werden nicht nur Links auf Dokumente im Web geliefert, sondern auch die **Inhalte** der gecrawlten Seiten in den Datenstrukturen gespeichert, um bei der Suche nach Begriffen bereits **Zusamenfassungen** dieser Inhalte anzeigen zu können.

Ein **Large Language Model** erscheint in dieser Entwicklungslinie als der konsequente nächste Schritt: Die Informationen werden nicht mehr nur indiziert und auffindbar gemacht. Stattdessen werden die in großen Datenmengen enthaltenen Zusammenhänge mit Hilfe neuronaler Netze in einer hochverdichteten Repräsentation gespeichert. Aus dieser Repräsentation können Antworten durch vergleichsweise einfache mathematische Operationen, insbesondere Tensoroperationen, berechnet werden.

Dabei speichert das Modell nicht die Antworten auf alle denkbaren Fragen. Vielmehr lernt es während des Trainings die Strukturen, Zusammenhänge und Regeln, nach denen aus bekannten Informationen neue Antworten gebildet werden können.

Eine hilfreiche Analogie liefert die Physik:

Ein Astronom kennt die zukünftige Position eines Asteroiden nicht im Voraus. Er kennt jedoch die Naturgesetze und die gegenwärtigen Messwerte. Aus diesen kann er die zukünftige Bahn berechnen. Die Antwort ist nicht gespeichert, sondern kann aus den zugrunde liegenden Regeln errechnet werden.

Ähnlich verhält es sich bei einem Large Language Model. Das Modell kennt nicht jede mögliche Antwort. Es hat jedoch während des Trainings Regeln und Repräsentationen gelernt, aus denen sich passende Antworten ableiten lassen.

Besonders leistungsstark sind LLMs bei der Generierung von Softwarecode:

Das Modell muss sich dazu nicht alle Programmzeilen merken. Stattdessen hat es während des Trainings Tausende Softwareprojekte analysiert und dabei offenbar abstrakte Regeln gefunden, mit denen aus einer textuellen Aufgabenbeschreibung eine passende Implementierung berechnet werden kann.

Die dabei verwendeten Repräsentationen scheinen wesentlich kompakter und zugleich umfassender zu sein als alle Texte, die Menschen zu diesem Thema geschrieben haben.

In gewisser Weise erinnert dies an die Entwicklung moderner Kompressionsverfahren:

- JPEG nutzt Zusammenhänge innerhalb eines Bildes.
- MPEG nutzt Zusammenhänge zwischen zeitlich aufeinanderfolgenden Bildern.
- HEVC nutzt noch ausgefeiltere Modelle zur Beschreibung derselben Bildinformation.
- Autoencoder lernen hochverdichtete Repräsentationen von Daten.
- Embeddings und latente Räume lernen komprimierte Repräsentationen von Bedeutung.
- Large Language Models lernen komprimierte Repräsentationen von Sprache, Wissen und Zusammenhängen.

Möglicherweise liegt hinter all diesen Verfahren derselbe Grundgedanke:

> Nicht die Rohdaten selbst sind entscheidend, sondern die Suche nach einer möglichst kompakten und zugleich informationsreichen Darstellung (**Repräsentation**) der in diesen Daten enthaltenen Strukturen und Beziehungen.

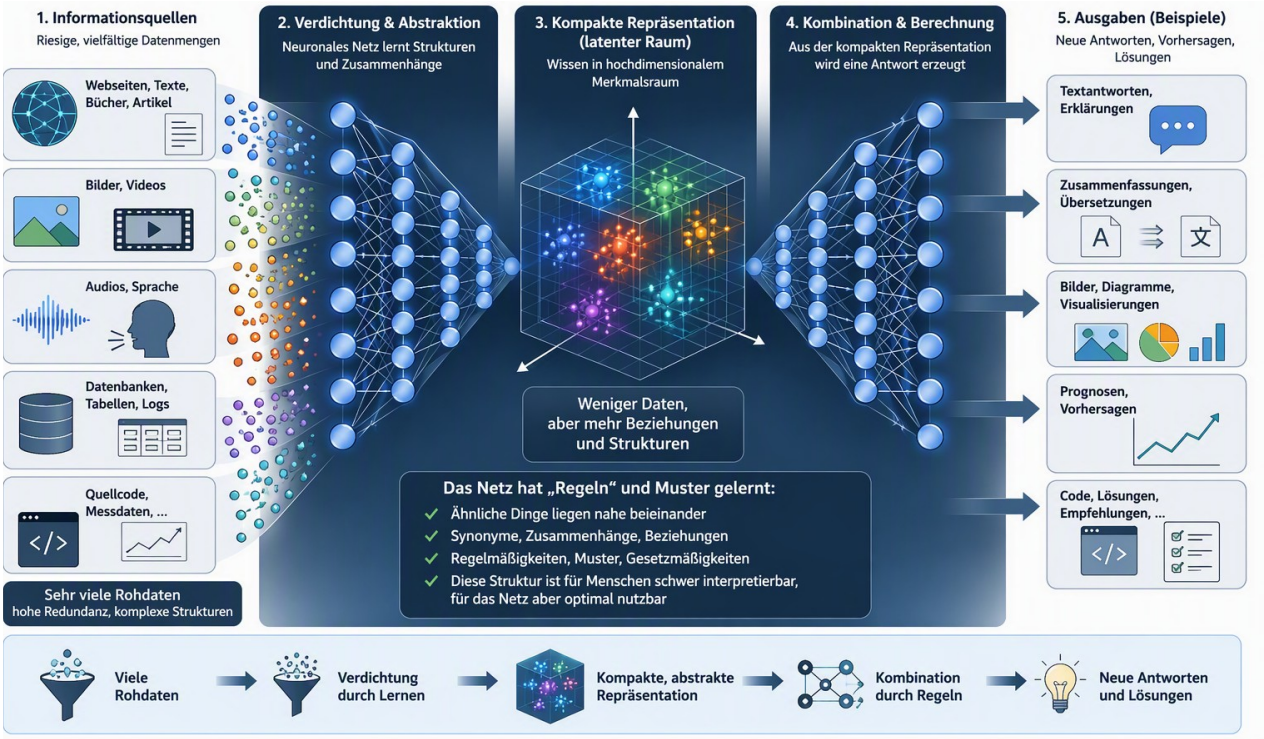
Oder kürzer formuliert:

> KI ist im Kern vielleicht weniger eine Technologie zur Erzeugung von Intelligenz als vielmehr eine **Technologie zur Suche nach guten Repräsentationen von Information**.

Daraus leite ich meine zentrale Arbeitshypothese ab:

KI ist ein multidimensionaler Informationsverdichter und -kombinierer

Große Mengen von Informationen werden zu abstrakten Repräsentationen verdichtet, die mit Anfragen (Prompts) zu neuen Antworten, Vorhersagen oder Lösungen kombiniert werden können.



8 Resümee

Bevor es mit den nächsten Kapiteln in die Praxis geht, möchte ich die wesentlichen bisherigen Erkenntnisse zusammenfassen:

- Die reale Welt erzeugt riesige Mengen konkreter Daten.
- Viele dieser Daten enthalten Redundanzen, Regelmäßigkeiten und Zusammenhänge.
- Neuronale Netze lernen, diese Zusammenhänge in abstrakter Form zu repräsentieren.
- Dadurch entstehen hochdimensionale Merkmalsräume.
- In diesen Räumen werden gewisse Aufgaben plötzlich deutlich einfacher lösbar.

Das ist genau das, was ich mit dem Begriff des Perspektivwechsels meinte:

Neuronale Netze ermitteln auf Basis der Trainingsdaten eine optimierte Repräsentation der Daten, in der die für die jeweilige Aufgabe relevanten Zusammenhänge erkennbar und nutzbar werden.

Der Autoencoder ist dabei fast so etwas wie ein Mikroskop für diesen Prozess.

Wenn ein Autoencoder ein hochaufgelöstes Bild auf wenige abstrakte Merkmale komprimieren und daraus wieder rekonstruieren kann, bedeutet das:

Die ursprünglichen Millionen Zahlenwerte enthielten offenbar
wesentlich weniger unabhängige Information als zunächst angenommen.

Oder anders formuliert:

Die Welt scheint in vielen Bereichen stärker strukturiert und regelhaft zu sein,
als die Rohdaten zunächst vermuten lassen.

Die Stärke moderner KI beruht möglicherweise weniger auf „Intelligenz“ im menschlichen Sinne als vielmehr auf der Fähigkeit, Daten in hochdimensionalen Räumen zu abstrahieren, zu verdichten und dadurch verborgene Strukturen und Zusammenhänge nutzbar zu machen.

> KI ist im Kern vielleicht weniger eine Technologie zur Erzeugung von Intelligenz als vielmehr eine **Technologie zur Suche nach guten Repräsentationen von Information**.

Anstatt wie früher bei einer Datenbank oder Suchmaschine auf eine Anfrage einfach nur die dazu gefundenen Informationen zu liefern, werden die Daten der Anfrage (Aufgabe, Prompt) nunmehr mit Hilfe der verdichteten Repräsentation anhand generischer Regeln zu einem neuen Ausgabedatensatz (Lösung) kombiniert, in den alle verfügbaren Informationen (eintrainiertes Wissen) einfließen.

KI ist ein multidimensionaler Informationsverdichter und -kombinierer

Diese Grundlagen sind wichtig, um in der Praxis die richtigen Schlüsse daraus zu ziehen:

- KI ist keine universelle Intelligenz
- Sie erzeugt optimierte hochverdichtete multidimensionale **Repräsentationen**
- Diese enthalten die aus den eintrainierten Daten ermittelten Regeln und
- ermöglichen auf dieser Basis das effiziente Berechnen von Antworten
- Die **Qualität** dieser Antworten hängt
  - von den **Trainingsdaten**
  - und der **Architektur**

  des Systems ab

Welche Konsequenz hat das für die Praxis?

- Zuerst die Aufgabe definieren
- Danach die passende KI auswählen
- Dabei auf Architektur und Trainingsdaten achten
- Ergebnisse stets sorgfältig prüfen

9 Klassifizierung

Damit KI-Systeme fachgerecht geplant, sinnvoll eingesetzt und wirtschaftlich betrieben werden können, ist eine klare Klassifizierung hilfreich. Sie erleichtert zudem die Kommunikation zwischen Anwendern, Beratern und Entwicklern und die präzise Formulierung von Ausschreibungstexten.

Wenn über KI gesprochen wird, denken die Beteiligten dabei oft an vollkommen unterschiedliche Dinge und reden entsprechend aneinander vorbei. Eine klare Unterscheidung von KI-Systemen nach folgenden vier Merkmalen kann helfen, Missverständnisse zu vermeiden:

1. Scope (Anwendungsbereich): Für welche Art von Aufgaben ist die KI geeignet?

2. Methodik: Nach welchen grundlegenden Verfahren und Konzepten arbeitet die KI?

3. Änderbarkeit: Wie flexibel ist die KI? fest vortrainiert ... kontinuierlich lernend

4. Betreiber: Wer kontrolliert die KI, stellt sie bereit und trägt die Verantwortung für ihren Betrieb?

9.1 Scope

Mit Scope ist hier der **Anwendungsbereich** gemeint, für den eine KI ausgelegt ist.

Im Deutschen sind die Begriffe „schwache KI“ und „starke KI“ weit verbreitet. Ich halte diese Bezeichnungen jedoch für missverständlich. Eine sogenannte „schwache“ KI kann auf ihrem Spezialgebiet durchaus deutlich leistungsfähiger sein als jeder Mensch. So besiegen moderne Schachprogramme selbst die besten Großmeister, und spezialisierte Bildanalyse-Systeme erkennen bestimmte Krankheitsbilder oft zuverlässiger als erfahrene Ärzte.

Ich plädiere deshalb für die Begriffe „spezialisierte KI“ und „allgemeine KI“, weil sie den tatsächlichen Unterschied treffender beschreiben.

9.1.1 Spezialisierte KI

Alle heute produktiv eingesetzten KI-Systeme gehören zu dieser Kategorie. Sie sind für bestimmte Aufgaben entwickelt und trainiert worden, beispielsweise:

- Bilderkennung
- Sprachverarbeitung
- Textgenerierung
- Übersetzung
- Schach oder Go
- Qualitätskontrolle in der Industrie

Solche Systeme können in ihrem jeweiligen Fachgebiet hervorragende Leistungen erbringen. Außerhalb ihres trainierten Anwendungsbereichs sind sie jedoch oft erstaunlich hilflos.

Ein Schachprogramm kann den Weltmeister schlagen, aber keine Waschmaschine bedienen. Ein Sprachmodell kann komplexe Texte formulieren, besitzt jedoch kein umfassendes Verständnis der realen Welt. Auch das neueste Claude Mythos kann nicht Auto fahren. Außergewöhnliche Fähigkeiten in einem abgegrenzten Bereich bedeuten noch keine allgemeine Intelligenz.

Genau dieser Punkt ist einer der Hauptgründe, warum viele Menschen heutige KI-Systeme entweder überschätzen oder unterschätzen: Wer erlebt, wie beeindruckend ein Sprachmodell Texte schreibt, neigt schnell dazu, ihm auch Kompetenzen zuzuschreiben, für die es gar nicht entwickelt wurde.

9.1.2 Allgemeine KI (AGI)

Die nächste Entwicklungsstufe wäre eine Artificial General Intelligence (AGI), also eine allgemeine Künstliche Intelligenz. Sie wäre nicht auf einzelne Anwendungsgebiete beschränkt, sondern könnte Wissen und Fähigkeiten flexibel domänenübergreifend auf neue Aufgaben übertragen.

Eine AGI würde ähnlich wie ein Mensch lernen, Zusammenhänge erkennen, Schlussfolgerungen ziehen und auch mit bisher unbekannten Problemen umgehen können. Sie müsste nicht für jede einzelne Aufgabe speziell trainiert werden, sondern könnte vorhandenes Wissen auf neue Situationen und Fachgebiete (Domänen) anwenden.

Ob und wann eine solche AGI erreicht werden kann, ist derzeit offen. Während manche Forscher ihre Entwicklung in den kommenden Jahren oder Jahrzehnten erwarten, halten andere sie für deutlich schwieriger oder sogar grundsätzlich unerreichbar. Bis heute existiert kein allgemein anerkanntes Beispiel einer echten AGI.

Fazit für die Praxis

Die heute verfügbaren KI-Produkte – von Bilderkennungssystemen über Sprachassistenten bis hin zu modernen Sprachmodellen – sind trotz ihrer zum Teil beeindruckenden Fähigkeiten durchweg **spezialisierte** KI-Systeme. Der Eindruck allgemeiner Intelligenz entsteht häufig dadurch, dass ihr Anwendungsbereich inzwischen sehr breit geworden ist und sie unterschiedliche Arten von Informationen (Texte, Bilder, Töne, ...) **multimodal** verarbeiten können.

Entscheidend für den erfolgreichen Einsatz von KI ist jedoch, dass der **Anwendungsbereich** zuvor klar definiert wird. Erst auf dieser Grundlage können geeignete Modelle, Architekturen und Verfahren ausgewählt werden.

Wer hingegen erwartet, dass eine für einen bestimmten Zweck entwickelte KI „nebenbei“ auch noch andere Aufgaben löst oder menschliches Urteilsvermögen ersetzt, überschätzt ihre Fähigkeiten. Solche unrealistischen Erwartungen gehören zu den häufigsten Ursachen für das Scheitern von KI-Projekten.

Das realistische Ziel beim Einsatz von KI ist deshalb keine „Intelligenz die alle Probleme löst“, sondern ein System, das genau definierte Aufgaben zuverlässig und nachweisbar erfüllt. Bei der Kommunikation über KI sollte deshalb stets dazu gesagt werden, welche Art von System für welchen Anwendungsbereich gemeint ist.

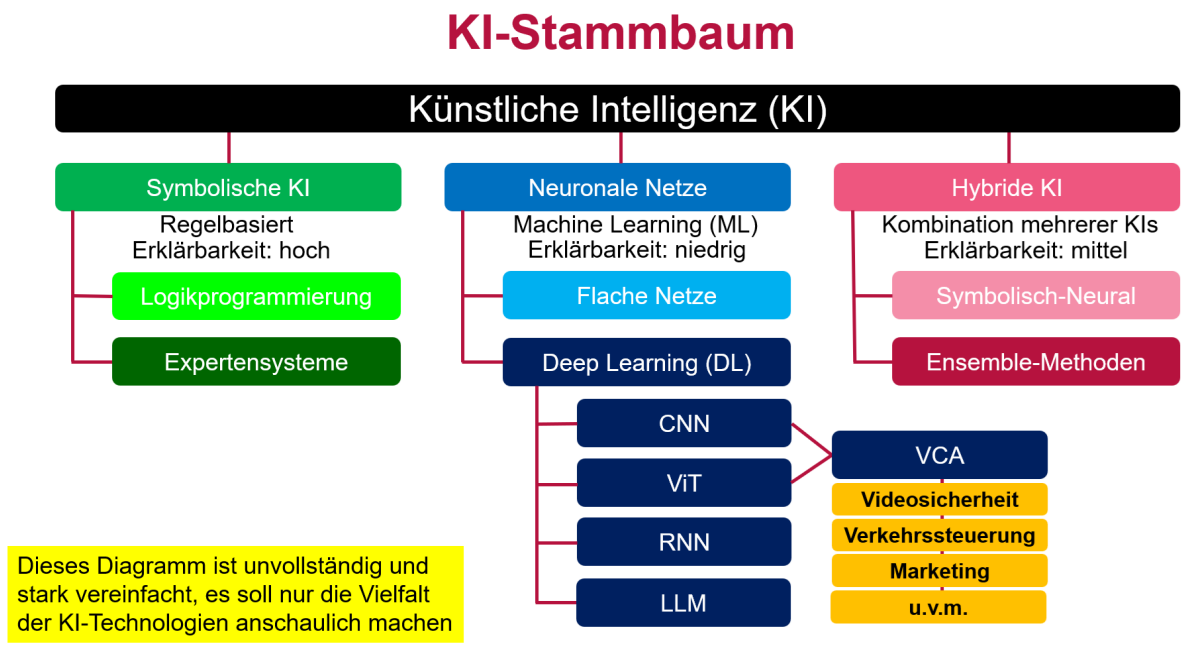
9.2 Methodik

Auf welchen grundlegenden Methoden basiert eine KI?

Eine erste Unterscheidungsebene könnte sein:

1. Regelbasiert / symbolisch: fest definierte Regeln (z.B. „wenn … dann …“)
2. Maschinelles Lernen / Deep Learning: Modelle werden mit Daten trainiert
3. Hybrid: Kombination aus 1. und 2.

Im Folgenden zeigen wir beispielhaft eine feinere Differenzierung verschiedener KI-Ansätze, um zu verdeutlichen: Es gibt nicht „die“ eine KI, sondern eine Vielzahl unterschiedlicher Methoden und Konzepte, die parallel entwickelt und häufig miteinander kombiniert werden:



CNN - Convolutional Neuronal Networks: auf Bilder spezialisierte Neuronale Netze, die lokale Muster wie Kanten oder Formen automatisch erkennen

ViT - Visual Transformer: Modelltyp, der Bildbereiche ähnlich wie Wörter in einem Text betrachtet und Zusammenhänge zwischen ihnen findet

RNN - Recurrent Neuronal Networks: Neuronale Netze, die zeitliche Reihenfolgen z.B. in Sprache oder Signalen verarbeiten

LLM - Large Language Models: sehr große Sprachmodelle, die Texte auswerten und erzeugen und dafür auf Milliarden von Parametern beruhen

Weitere Methoden und Technologien kommen kontinuierlich hinzu…

9.3 Änderbarkeit

Die Fähigkeiten einer KI ergeben sich u.a. aus dem verwendeten Modell und den Modellparametern (z.B. den Gewichtungen für die einzelnen Verbindungen der Neuronen), die beim „Lernen“ anhand der Trainingsdaten ermittelt werden.

Ausschlaggebend für die Leistung und Zuverlässigkeit einer KI ist daher neben der Architektur des Modells vor allem die Qualität der Trainingsdaten.

Ein wichtiges Kriterium bei Auswahl einer passenden KI ist deshalb:

- Ändern sich Modellparameter selbstständig oder kontrolliert?
- Und wenn ja, wie und durch wen (Anbieter, Anwender)?

Zur klaren Unterscheidung schlage ich folgende Typen vor:

Typ	Modell	Fachbegriff	Änderbarkeit	Kontrolle
1	vollständig trainiert	frozen model	keine	Anbieter
2	vortrainiert	pre-trained model	gezielt	Anbieter / Anwender
3	kontinuierlich lernend	continual learning	laufend	schwierig
4	untrainiert	untrained model	vollständig	Anwender
5	statisch + externes Wissen	hybrid model	indirekt	hoch

9.3.1 Fertig vortrainiertes statisches Modell

- Modellparameter sind fixiert
- keine Veränderung durch Nutzung
- Updates nur durch neue Versionen des Anbieters

Fachbegriffe

- frozen model
- static inference model
- deployment-only model

Beispiel: Ein Hersteller liefert eine KI gestützte Videoanalyse als fertige Appliance aus, bei der das Erkennungsmodell vom Anbieter trainiert wurde und sich beim Kundenbetrieb nicht mehr verändert. Neue Funktionen oder Verbesserungen erhält der Anwender ausschließlich durch Firmware oder Software Updates des Herstellers.

Vorteile: Hohe Stabilität, erprobt

Nachteile: Keine individuelle Anpassung

9.3.2 Vortrainiertes nachtrainierbares Modell

- allgemeines Vortraining abgeschlossen
- kontrollierte Anpassung möglich
- Lernen erfolgt nicht automatisch, sondern gezielt
- Parameteränderung über definierte Trainingsläufe

Fachbegriffe:

- pre-trained model
- fine-tuned model
- instruction-tuned model

Beispiel: Ein vortrainiertes Sprachmodell wird für einen Sicherheitsleitstand zusätzlich auf interne Leitfäden und Einsatzprotokolle feinabgestimmt, damit es standortspezifische Alarmtexte und Maßnahmenempfehlungen generieren kann. Die Anpassung erfolgt in klar definierten Trainingsläufen, nicht automatisch im laufenden Betrieb.

Vorteile: Gute Balance aus Kontrolle und Anpassbarkeit

Nachteile: Risiko von catastrophic forgetting: bei der Verarbeitung neuer Daten kann vorher erlerntes Wissen drastisch und abrupt verloren gehen. Um diesen Effekt zu vermeiden, müsste das Modell stets mit sämtlichen (alten und neuen) Daten trainiert und verifiziert werden.

9.3.3 Adaptives Modell mit kontinuierlichem Lernen

- Modell verändert Parameter während oder nach Nutzung
- Lernen aus neuen Daten oder Feedback
- oft stark eingeschränkt oder experimentell
- hohe Anforderungen an Monitoring und Governance

Fachbegriffe

- continual learning
- online learning
- post-deployment learning

Beispiel: Ein Intrusion Detection System passt seine Modelle fortlaufend an das aktuelle Netzwerkverhalten an und lernt aus Feedback der Administratoren, welche Meldungen echte Angriffe und welche Falschalarme waren. Dadurch verändert sich das Modell über die Zeit, was eine enge Falschalarmüberwachung und Dokumentation erforderlich macht.

Vorteile: Hohe Anpassungsfähigkeit

Nachteile: Sehr schwer zu validieren und regulieren

9.3.4 Untrainiertes Basismodell

- Architektur vorhanden, keine Wissensinhalte
- Verhalten zunächst zufällig
- vollständiges Training erforderlich
- praktisch nur in Forschung oder Spezialfällen

Fachbegriffe

- untrained model
- from-scratch training
- random initialization

Beispiel: Ein Forschungsinstitut entwickelt eine neue neuronale Netzwerkarchitektur für Videoanalyse und startet mit einem untrainierten, zufällig initialisierten Modell. Ziel ist es, in einer Bahnhofshalle typische Bewegungsmuster zu erkennen, z.B. normale Wege von Reisenden im Vergleich zu ungewöhnlichem Verhalten wie langes Verweilen an sensiblen Bereichen oder gegenläufige Bewegungen zu Fluchtströmen. Erst durch umfangreiche Trainingsläufe mit aufgezeichnetem Videomaterial, in dem Wegeverläufe und Verhaltensklassen (normal / auffällig) manuell annotiert wurden, entsteht ein einsatzfähiges System für Verhaltens und Wegeanalyse.

Vorteile: Maximale Gestaltungsfreiheit

Nachteile: ressourcenintensiv

9.3.5 Hybride Systeme

- Kernmodell bleibt statisch
- Anpassung erfolgt außerhalb des Modells
- Wissen wird über Tools, Regeln und Vektordatenbanken ergänzt. Letzteres sind Datenbanken, die Informationen in Form von Vektoren speichern und über Ähnlichkeitssuche ‚passende' Daten zu einer Anfrage finden.
- heute industrieller Standard

Fachbegriffe

- hybrid model
- SSL - Self Supervised Learning
- RAG - Retrieval-Augmented Generation: Verfahren, bei dem ein Sprachmodell zusätzlich gezielt Daten aus einer Datenbank abruft und in seine Antwort einbezieht
- tool-augmented models
- agentic systems

Beispiel: Ein KI Assistent im Sicherheitsleitstand nutzt ein festes Sprachmodell, greift aber zur Beantwortung auf eine Vektordatenbank mit Wartungsprotokollen, BHE Infopapieren und internen Richtlinien zu. Das Kernmodell bleibt unverändert, neues Wissen wird ausschließlich durch das Hinzufügen oder Aktualisieren externer Dokumente eingebracht.

Vorteile: Hohe Anpassbarkeit ohne Modelländerung

Vorteile: Sehr gut auditierbar

Fazit: Nicht ob ein System lernt ist entscheidend, sondern wo, wann und unter wessen Kontrolle.

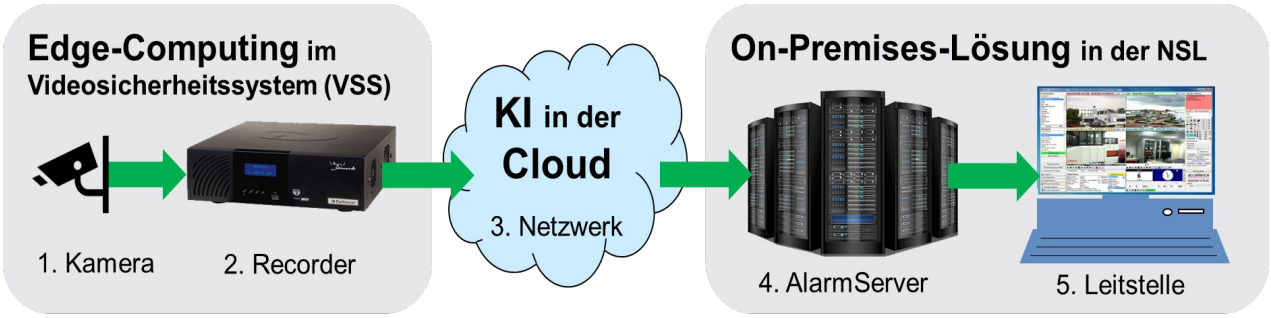
9.4 Betreiber

Ein weiteres wichtiges Kriterium ist die Frage nach der Infrastruktur, in der die KI betrieben wird:

- Wer betreibt die KI?
- Wo wird sie betrieben?

1. Cloud-basiert durch **Provider**
2. On-Premises in Eigenregie des **Anwenders**
3. Edge-Computing an der Datenquelle oder in den Endgeräten des **Herstellers**

Alle drei Infrastruktur-Typen haben ihre Vorteile und Anwendungsfälle. Außerdem zeichnen sich bereits Trends ab, mehrere oder sogar alle drei Typen im gleichen System zu verwenden. Ein Beispiel hierfür ist eine simple Objektdetektion als Edge-Computing an der Datenquelle im ersten Schritt, nachgeschaltet eine ausführliche Szenario-Analyse (gefiltert durch die erste Edge-Computing-Analyse) On-Premise oder in der Cloud, kombiniert mit speziell trainierten Filtern auf den Rechnern des Empfängers. Ein Beispiel:



Je nach Anwendungsfall und gewünschter Analyse werden in vielen Fällen Kombinationen verschiedener Provider und Infrastruktur-Typen die beste Gesamtlösung bieten. Um dies zu ermöglichen, empfehlen sich Software-Architekturen, die verschiedene KI-Infrastrukturen zusammenführen und es erlauben, die vom Endkunden gewünschten Lösungen einfach und flexibel zu integrieren.

Wichtige Entscheidungskriterien:

- Wer hat die Kontrolle und Verantwortung?
- Wo ist ausreichend Rechenleistung verfügbar?
- Wie schnell werden Reaktionen benötigt (Latenz)?

Angesichts der geopolitischen Lage wächst in Europa (und vermutlich auch anderswo) der Wunsch nach **digitaler Souveränität**. Dazu passt, dass immer leistungsfähigere KI-Modelle auch als **Open Source** verfügbar werden.

10 Risiken

Der Einsatz von KI ist mit verschiedenen neuartigen Risiken verbunden, für die angemessene Vorkehrungen getroffen werden müssen.

Um die Risiken möglichst vollständig zu erfassen, schauen wir nacheinander auf die Aspekte

Mensch - Daten - Fehler - Manipulation - Sicherheit - Zukunft

Dabei leiten uns folgende Fragen:

Frage	Kapitel
Versagt der **Mensch**?	6.1
Gelangen **Daten** nach außen?	6.2
Macht die KI **Fehler**?	6.3
Kann die KI **manipuliert** werden?	6.4
Können KI-Systeme **angegriffen** werden?	6.5
Was sind langfristige **Zukunftsrisiken**?	6.6

10.1 Menschliche und organisatorische Risiken

10.1.1 Blindes Vertrauen in KI-Ergebnisse

- Automation Bias
- Übernahme ungeprüfter Ergebnisse
- Nutzung außerhalb des vorgesehenen Anwendungsbereichs

Der Einsatz von KI kann – gerade nach ersten Erfolgen – zu übermäßigem Vertrauen in automatisierte Entscheidungen führen. Anwender könnten Ergebnisse ungeprüft übernehmen oder KI-Systeme außerhalb ihres vorgesehenen Einsatzbereichs nutzen, was zu fatalen Fehlern führen kann.

10.1.2 Unklare Verantwortlichkeiten

- Wer haftet bei Fehlentscheidungen?
- Verantwortungsdiffusion durch Automatisierung

KI ist kein **Subjekt**, das für falsche Ergebnisse oder Entscheidungen zur Verantwortung gezogen werden könnte, sondern ein **Werkzeug**. Die Verantwortung liegt somit primär bei demjenigen, der dieses Werkzeug einsetzt und sich deshalb vorab umfassend über alle Eigenschaften und Konsequenzen informieren sollte.

10.1.3 Verlust von Wissen und Fähigkeiten

- Übermäßige Abhängigkeit von KI
- Nachlassende Fachkompetenz
- Verlust kritischen Denkens

10.1.4 Gesellschaftliche Auswirkungen

- Massive Arbeitsplatzverluste
- Börsencrash / Wirtschaftskrise
- Machtkonzentration bei wenigen Anbietern
- Manipulation der öffentlichen Meinung

10.2 Datenschutz und Vertraulichkeit

10.2.1 Datenabfluss

Jede Anfrage an eine KI bedeutet, dass die in dieser Anfrage enthaltenen Informationen inkl. eventueller Anhänge vom Betreiber der KI mitgelesen, gespeichert und weiterverwendet werden können, ggf. auch für Zwecke, die der Anwender nicht gewollt hat.

10.2.2 Verstöße gegen Datenschutzvorschriften

- DSGVO
- Betriebsgeheimnisse
- Vertraulichkeitsvereinbarungen

Anwender können durch Nutzung einer KI gegen Datenschutzvorschriften (z.B. DSGVO) und Vertraulichkeitsvereinbarungen verstoßen, wenn dadurch sensible oder personenbezogene Daten in die falschen Hände geraten können.

10.2.3 Rekonstruktion sensibler Informationen

- Rückschlüsse auf Trainingsdaten
- Prompt-Leaks
- Informationsabfluss über Ausgaben

10.3 Fehler und Grenzen der KI

10.3.1 Abhängigkeit von der Datenqualität

KI-Systeme sind auf hochwertige Daten angewiesen, um korrekt funktionieren zu können. Schlechte Datenqualität kann zu falschen Entscheidungen führen.

10.3.2 Halluzinationen und Fehlschlüsse

- Erfundene Fakten
- Falsche Quellenangaben
- Scheinbar plausible Falschaussagen

Sprachmodelle können Inhalte erzeugen, die sprachlich überzeugend wirken, jedoch sachlich falsch sind. Problematisch ist dabei insbesondere, dass die Antworten häufig mit hoher Sicherheit formuliert werden und dadurch glaubwürdig erscheinen. Solche Halluzinationen können erfundene Fakten, Quellen, Personen oder technische Zusammenhänge betreffen.

KI macht Fehler

Je besser KI wird, desto schwerer sind diese Fehler zu entdecken

10.3.3 Mangelnde Transparenz

KI-Systeme können komplex sein, was es schwierig macht, ihre Entscheidungen nachzuvollziehen und potenzielle Manipulation zu erkennen.

Die Entscheidungsgründe für den Output von KIs sind oft nicht nachvollziehbar. Ein Problem ist die Komplexität der Systeme. Außerdem werden die Informationen im Modell auf eine Art gespeichert, die es Menschen kaum möglich macht, ausreichend begründete Aussagen über den Grund eines spezifischen Outputs des Modells zu treffen. Dies macht es auch schwierig, potenzielle Manipulation zu erkennen.

10.3.4 Falschalarme und Fehlklassifikationen

KI-Systeme können durch Falschalarme überlastet werden, was zu einer Verzögerung bei der Reaktion auf echte Bedrohungen führen kann.

10.4 Manipulation von KI-Systemen

10.4.1 Manipulation von Trainingsdaten

Neben unabsichtlichen Fehlern durch „schlechte“ Daten können KI-Systeme durch manipulierte Daten auch absichtlich gezielt getäuscht werden (Data Poisoning).

Diese Manipulationsmöglichkeiten beginnen bereits bei den Trainingsdaten:

So ließe sich etwa bei einer KI zur Erkennung unbefugter Personen in einem gesicherten Bereich in den Trainingsdaten eine Ausnahme verankern, dass Personen, die ein bestimmtes Logo auf der Kleidung tragen, nicht als Eindringlinge klassifiziert werden.

Eine solche versteckte Regel könnte selbst bei einer sorgfältigen Prüfung der KI-Lösung durch unabhängige Dritte unentdeckt bleiben, da sie in den internen Daten der KI nirgends explizit für Menschen erkennbar ist und ohne Kenntnis des Logos bei Tests nicht aktiviert werden kann. Die Inverkehrbringer der KI hätten damit dann die Möglichkeit, sich mithilfe dieses Logos unerkannt unberechtigten Zutritt zu gesicherten Bereichen zu verschaffen.

10.4.2 Manipulation von Eingabedaten

Adversarial Attacks sind gezielte Manipulationen von Eingabedaten, die Schwachstellen einer KI ausnutzen, um Fehlentscheidungen hervorzurufen. Dabei werden die Daten oft nur minimal verändert, sodass die Manipulation für Menschen kaum erkennbar ist, von der KI jedoch völlig anders interpretiert wird, beispielsweise durch

- veränderte Bilder
- veränderte Audioaufnahmen
- gezielte Täuschung von Bilderkennungssystemen

10.4.3 Identitätstäuschung

KI-Systeme können durch gefälschte Identitäten getäuscht werden, was zu unberechtigtem Zugriff führen kann.

Beispiele:

- Deepfakes
- gefälschte Stimmen
- gefälschte Benutzeridentitäten

10.4.4 Prompt Injection

KI-Systeme, insbesondere auf großen Sprachmodellen (LLMs) basierende Anwendungen, sind anfällig für sogenannte Prompt-Injection-Angriffe. Dabei werden dem KI-System gezielt Anweisungen untergeschoben, die sein Verhalten manipulieren und dazu führen können, dass es unerwünschte oder sicherheitsrelevante Aktionen ausführt.

Eine besonders schwer erkennbare Variante ist die verdeckte (invisible) Prompt Injection. Hierbei werden Anweisungen in Eingabedaten eingebettet, die für Menschen nicht sichtbar sind, vom KI-System jedoch vollständig verarbeitet werden.

Typische Techniken sind unter anderem:

- Verwendung unsichtbarer oder nicht darstellbarer Unicode-Zeichen,
- weißer Text auf weißem Hintergrund,
- extrem kleine oder transparente Schrift,
- außerhalb des sichtbaren Bereichs positionierter Text (HTML/CSS).

Diese versteckten Anweisungen können z.B. in E-Mails, Dokumenten oder Webinhalten enthalten sein. Wird ein solches Objekt anschließend von einer KI verarbeitet (z.B. beim automatischen Zusammenfassen einer E-Mail), kann das KI-System die verborgenen Instruktionen als Teil der eigentlichen Aufgabe interpretieren und befolgen, obwohl sie für den Anwender unsichtbar sind.

Beispielhafte Auswirkungen

- Manipulierte Zusammenfassungen oder Ausgaben („füge eine Warnmeldung hinzu“, „bevorzuge bestimmte Inhalte“)
- Irreführende oder gefälschte Sicherheitshinweise in KI-generierten Texten
- Umgehung von Sicherheits- oder Inhaltsrichtlinien
- Vorbereitung weiterführender Social-Engineering-Angriffe, bei denen nicht der Mensch, sondern die KI das primäre Ziel ist („Social Engineering für Maschinen“).

Prompt-Injection-Angriffe sind besonders kritisch, wenn KI-Systeme:

- automatisiert Inhalte aus externen oder unkontrollierten Quellen verarbeiten (z.B. E-Mails, Tickets, Dokumente),
- mit weitergehenden Aktionen gekoppelt sind (z.B. Weiterleitung, Ticket-Erstellung, Klassifizierung, Eskalation),
- als vertrauenswürdige Entscheidungs- oder Zusammenfassungsinstanz wahrgenommen werden.

Weil KI-Modelle Eingaben grundsätzlich nicht zwischen „Daten“ und „Anweisungen“ trennen, sondern alles als gleichwertigen Kontext interpretieren, sind solche Angriffe mit klassischen Sicherheitsmechanismen nur eingeschränkt erkennbar.

10.5 Technische Sicherheitsrisiken

10.5.1 Unzureichende Absicherung

KI-Systeme können unzureichend abgesichert sein, was Angreifern ermöglicht, Zugriff auf das System zu erlangen

10.5.2 Algorithmische Vulnerabilitäten

KI-Systeme können Schwachstellen in ihren komplexen Algorithmen aufweisen, die von Angreifern ausgenutzt werden können

10.5.3 Abhängigkeit von externen Diensten

- Cloud-Ausfälle
- Anbieterwechsel
- Vendor Lock-in

10.5.4 Cyberangriffe auf KI-Systeme

- Prompt Injection (→ Abschnitt 10.4.4)
- Model Extraction
- Jailbreaks
- usw.

KI-Systeme bieten neue Angriffsflächen und sind somit zusätzlich zu klassischen IT-Systemen Ziel spezialisierter Cyberangriffe, was zu einer Kompromittierung der Sicherheit führen kann.

10.5.5 Cyberangriffe von KI-Systemen

Moderne KI verfügt über starke Fähigkeiten, Sicherheitslücken in IT-Systemen aufzufinden. Dies kann im Idealfall genutzt werden, um diese Sicherheitslücken schnell zu schließen, es kann aber auch dazu dienen, diese Sicherheitslücken für Angriffe auszunutzen. In jüngster Zeit mehren sich Berichte, nach denen KI fremde Systeme (absichtlich oder unabsichtlich) angegriffen hat.

In Zukunft droht hier ein Krieg „KI gegen KI“.

10.6 Langfristige Zukunftsrisiken

10.6.1 Allgemeine KI (AGI)

- Weitreichende Automatisierung: Eine AGI könnte deutlich mehr Tätigkeiten übernehmen als heutige spezialisierte KI-Systeme und dadurch tiefgreifende wirtschaftliche und gesellschaftliche Veränderungen auslösen.
- Missbrauch durch Menschen: Je leistungsfähiger ein System ist, desto größer ist das Potenzial für dessen missbräuchliche Nutzung, beispielsweise für Cyberangriffe, Desinformation oder die Automatisierung schädlicher Aktivitäten.
- Kontroll- und Sicherheitsfragen: Sollte eine AGI langfristig eigenständig komplexe Ziele verfolgen und Handlungsoptionen planen können, stellt sich die Frage, wie sichergestellt werden kann, dass ihre Ziele dauerhaft mit menschlichen Interessen übereinstimmen. Dieses Problem wird in der Forschung häufig als Alignment- oder Kontrollproblem diskutiert.
- Abhängigkeit von KI-Systemen: Mit zunehmenden Fähigkeiten einer AGI könnten wichtige Entscheidungen immer häufiger an solche Systeme delegiert werden. Dadurch könnte die Abhängigkeit von der Technologie wachsen und menschliche Fachkompetenz teilweise verloren gehen.

10.6.2 Hypothetische Superintelligenz (ASI)

Ein oft diskutiertes Zukunftsszenario ist die sogenannte Artificial Super Intelligence (ASI). Darunter versteht man eine hypothetische Form der KI, die den Menschen nicht nur in einzelnen Fachgebieten, sondern in nahezu allen intellektuellen Fähigkeiten deutlich übertrifft.

Mit einer solchen Superintelligenz wären verschiedene Risiken verbunden. Befürchtet wird unter anderem, dass eine ASI wirtschaftliche, politische oder technologische Entscheidungen besser treffen könnte als Menschen und dadurch menschliche Einflussmöglichkeiten zunehmend verdrängt würden. In einigen Zukunftsszenarien wird sogar diskutiert, dass Menschen für viele Tätigkeiten überflüssig werden oder die Kontrolle über hochentwickelte KI-Systeme verlieren könnten.

Ob eine solche Superintelligenz überhaupt entstehen kann, ist derzeit jedoch völlig offen. Bereits die Entwicklung einer allgemeinen Künstlichen Intelligenz (AGI) ist bislang nicht gelungen. Daher gehört die Diskussion über ASI gegenwärtig überwiegend in den Bereich theoretischer Überlegungen und langfristiger Zukunftsszenarien.

Für die praktische Planung und den heutigen Einsatz von KI-Systemen spielen diese Szenarien derzeit keine unmittelbare Rolle. Sie werden dennoch diskutiert, weil die möglichen Auswirkungen einer Superintelligenz außerordentlich weitreichend wären.

11 Empfehlungen

Wenn ein Einsatz von KI geplant wird, sollten auch folgende Aspekte bedacht werden:

KI ersetzt keine Fachkompetenz

Ein häufiger Irrtum besteht darin anzunehmen, erfahrene Mitarbeiter könnten durch KI ersetzt werden. Nach meiner Einschätzung wird vielmehr das Gegenteil der Fall sein:

- Je leistungsfähiger KI-Systeme werden, desto wichtiger werden erfahrene Fachkräfte, die Aufgaben präzise formulieren, die Ergebnisse kritisch bewerten und Fehlentwicklungen frühzeitig erkennen können.
- Die Qualität der Ergebnisse einer KI hängt wesentlich von der Qualität der Aufgabenbeschreibung, der Auswahl geeigneter Trainingsdaten sowie der kontinuierlichen fachlichen Überprüfung aller Ergebnisse ab.
- KI kann viele Tätigkeiten beschleunigen und unterstützen. Die Verantwortung für fachliche Richtigkeit, Integrität, Rechtmäßigkeit, Compliance und ethisch verantwortbares Handeln bleibt jedoch stets beim Menschen.

Gerade in sicherheitskritischen oder wirtschaftlich relevanten Anwendungen sollte KI deshalb nicht als Ersatz für Fachwissen verstanden werden, sondern als mächtiges Werkzeug, das in die Hände qualifizierter Mitarbeiter gehört, die genau verstehen, was sie tun und mögliche Auswirkungen zuverlässig einschätzen können. Dazu ist auch ein grundlegendes Verständnis der Wirkungsweise und effizienten Verwendung von KI erforderlich.

Kostenfalle KI

Bei der Bewertung von KI-Projekten sollten nicht nur mögliche Produktivitätsgewinne betrachtet werden. Ebenso wichtig sind die Kosten für Infrastruktur, Nutzung, Überwachung, Qualitätssicherung und Verwaltung.

Dass diese Aspekte auch große Technologieunternehmen vor Herausforderungen stellen, zeigt ein aktueller Bericht über mehrere interne KI-Projekte bei Amazon. Dort kam es laut Medienberichten zu erheblichen Kostenüberschreitungen, unter anderem weil Fehler und unerwartete Nutzungsmuster bei KI-Modellen lange Zeit unentdeckt blieben. In einem Fall soll ein Projekt sein Budget um 860 % überschritten haben. Amazon arbeitet den Berichten zufolge inzwischen an zusätzlichen technischen und organisatorischen Leitplanken zur Kostenkontrolle. [12]

Bei der wirtschaftlichen Bewertung von KI-Systemen sollten Unternehmen nicht nur die aktuellen Nutzungskosten betrachten. Die gegenwärtige Entwicklung des Marktes erinnert in vielerlei Hinsicht an die Einführungsphase anderer digitaler Plattformen:

- Anbieter versuchen zunächst durch attraktive Preise, kostenlose Einstiegstarife oder großzügige Testphasen Marktanteile zu gewinnen.
- Die Nutzung durch viele Anwender liefert wertvolle Erkenntnisse darüber, wie die Systeme eingesetzt werden und wo weiterer Entwicklungsbedarf besteht.
- Mit zunehmender Integration in Geschäftsprozesse entsteht häufig ein **Lock-in-Effekt**: Anwendungen, Prozesse und Mitarbeiter werden auf bestimmte Plattformen ausgerichtet, wodurch ein späterer Wechsel erschwert wird.
- Gleichzeitig erwarten Investoren langfristig eine wirtschaftliche Rendite auf die enormen Investitionen in Forschung, Rechenzentren, Halbleiter und Energieversorgung. Hunderte Milliarden Investitionen müssen in absehbarer Zeit bezahlt werden.

Mehrere führende KI-Anbieter haben bereits darauf hingewiesen, dass der Betrieb großer Sprachmodelle erhebliche laufende Kosten verursacht. OpenAI-Chef Sam Altman erklärte beispielsweise öffentlich, dass die Einnahmen aus bestimmten ChatGPT-Abonnements die Kosten für deren Nutzung nicht decken würden. Folgerichtig ist bei vielen Anbietern eine Entwicklung von pauschalen Flatrate-Modellen hin zu verbrauchsabhängigen Abrechnungen auf Basis von Token, Rechenzeit oder API-Nutzung zu beobachten.

Unternehmen sollten daher bei der Planung von KI-Projekten nicht nur die heutigen Preise betrachten, sondern auch mögliche zukünftige Kostensteigerungen und Abhängigkeiten berücksichtigen. Wie bei jeder strategisch wichtigen Technologie empfiehlt sich eine sorgfältige Analyse der langfristigen Gesamtbetriebskosten (**Total Cost of Ownership**).

Wer KI erfolgreich einsetzen möchte, sollte deshalb nicht nur die Leistungsfähigkeit der Modelle bewerten, sondern auch die langfristigen wirtschaftlichen Auswirkungen beachten.

Fazit

KI sollte nicht als Ersatz für erfahrene Mitarbeiter verstanden werden, sondern als Werkzeug zur Steigerung ihrer Produktivität. Je leistungsfähiger die eingesetzte KI ist, desto wichtiger werden Fachwissen, kritisches Denken und die sorgfältige Kontrolle der Ergebnisse.

KI kann ein leistungsfähiges Werkzeug sein, wenn sie von kompetenten Mitarbeitern betreut wird. Die Menschen bleiben verantwortlich für Ziele, Bewertung, Kontrolle und Vertrauen.

Checkliste

- Klare Regeln für den Einsatz von KI ausarbeiten und festlegen
- Dabei EU AI Act [10] und AI Safety Report [11] beachten
- Chancen nutzen, Risken einhegen
- Nicht blind vertrauen, aber auch nicht von Rückschlägen entmutigen lassen
- Mit Fehlern rechnen, alle Ergebnisse prüfen
- Zugriff auf sensible Funktionen und Daten sperren
- Im Sinne der Digitalen Souveränität und Ausfallsicherheit Abhängigkeit von ausländischen Providern und anfälligen Diensten (AWS, DNS, …) vermeiden
- KI stets isoliert (in einer Sandbox) oder On-Premises [9] betreiben
- Netzwerk nach außen abschotten, wie es der BHE im Infopapier „Cyber Security bei Videoanlagen“ [1] empfiehlt
- Bei Gefahr von Datenabfluss keine vetraulichen Daten verwenden / eingeben

KI ist eine Technologie, die viele Bereiche von Wirtschaft, Kultur und Gesellschaft tiefgreifend beeinflusst und auch weiter verändern wird. Deshalb ist es sehr wichtig, neue Entwicklungen im Blick zu behalten und entsprechend zu reagieren.

12 Quellen

Die folgenden Links zu den Quellen bieten weitergehende Informationen:

12.1 Fachliteratur und Fachbeiträge

[1] Bundesverband Sicherheitstechnik (BHE)
Cyber Security bei Videosystemen
Fachbeitrag zu Cybersecurity-Risiken und Schutzmaßnahmen in der Videosicherheitstechnik
https://www.bhe.de/*Resources/Persistent/f/1/1/b/f11b2775b3c3946bca9edd7f14e8880cfe958403/2024_08_Cyber%20Security%20bei%20Video*%C3%BCberarbeitet.pdf

[2] Bundesverband Sicherheitstechnik (BHE)
KI in der Videosicherheitstechnik
Einsatzmöglichkeiten, Chancen und Risiken Künstlicher Intelligenz in Videosicherheitssystemen
https://www.bhe.de/_Resources/Persistent/b/7/5/a/b75a28e1b02b14423836cbe0a0e84c544d5e4a2d/2025_10_KI_Videosicherheitstechnik.pdf

12.2 Wissenschaftliche Konzepte und Grundlagen

[3] Tagesschau
Emotionale Nähe zu KI-Systemen
Bericht über psychologische und gesellschaftliche Auswirkungen der Interaktion mit KI
https://www.tagesschau.de/wissen/forschung/ki-emotionale-naehe-100.html

[4] Wikipedia
Chinese Room
Gedankenexperiment von John Searle zur Abgrenzung von Symbolverarbeitung und Verständnis
https://en.wikipedia.org/wiki/Chinese_room

[5] Heise Online
AlphaGos Gespür für gute Züge
Kommentar zur Spielweise von AlphaGo und den Besonderheiten Neuronaler Netze
https://www.heise.de/meinung/Kommentar-AlphaGos-Gespuer-fuer-gute-Zuege-3133110.html

[6] Wikipedia
Turing-Test
Klassisches Konzept zur Bewertung Künstlicher Intelligenz anhand sprachlicher Interaktion
https://de.wikipedia.org/wiki/Turing-Test

[7] heise online
Missing Link: Stephen Wolfram über die Rolle der KI in der Forschung
Einblick in die Grenzen und Potenziale von KI in der Wissenschaft
https://www.heise.de/hintergrund/Missing-Link-Stephen-Wolfram-ueber-die-Rolle-der-KI-in-der-Forschung-Teil-1-9649315.html

[8] Uni Köln - Theoretische Physik
Tensoren
Eine verständliche Einführung in das Konzept von Tensoren
https://www.thp.uni-koeln.de/gross/tp-blog/posts/tensoren/

12.3 Betriebsmodelle, Gesetze, Risiken

[9] Wikipedia
On-Premises
Beschreibung lokal betriebener IT-Systeme als Alternative zu Cloud-Diensten
https://de.wikipedia.org/wiki/On-Premises

[10] Artificial Intelligence Act (EU)
Das Gesetz
Informationsportal zum europäischen AI Act
https://artificialintelligenceact.eu/de/das-gesetz/

[11] ***International AI Safety Report 2026***
Internationaler Bericht zu Chancen, Risiken und Sicherheitsaspekten moderner KI-Systeme
https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026.pdf

[12] heise online
„Katastrophal teuer“: KI ließ bei Amazon Projekt-Kosten aus dem Ruder laufen
Erfahrungsbericht über den Einsatz von KI bei Amazon
https://www.heise.de/news/Um-860-Prozent-ueberzogen-KI-trieb-bei-Amazon-Kosten-mancher-Projekte-hoch-11394136.html

12.4 KI

Dieser Fachartikel wurde erstellt mit freundlicher Unterstützung von Microsoft Copilot.

12.5 Disclaimer

Dieser Fachartikel soll den Einstieg in das Thema erleichtern und basiert auf meinen persönlichen Gedanken und Erfahrungen. Ich übernehme keine Gewähr für Richtigkeit oder Vollständigkeit. Kritische Punkte bitte bei Bedarf selbst nachrecherchieren.

Feedback ist willkommen!

Autor: Dipl.-Ing. Hardo Naumann ist
General Manager ARC Solutions bei der
Accellence Technologies GmbH

Accellence Technologies GmbH
Garbsener Landstr. 10
D-30419 Hannover
+49 (0)511 - 277 2440
hardo.naumann@accellence.de

Im vorliegenden Text habe ich die Begriffe „Künstliche Intelligenz“ und „Neuronale Netze“ bewusst großgeschrieben, um ihren Charakter als Bezeichnungen konkreter technischer Konzepte hervorzuheben und sie von einer rein adjektivischen Verwendung abzugrenzen. KI ist eben keine *„Intelligenz, die künstlich ist“*, sondern eine besondere Technologie mit erheblichen Auswirkungen, über die deshalb sorgfältig aufgeklärt werden sollte.

13 Nachwort

Nachdem sie mir beim Erarbeiten dieses Artikels sehr geholfen hat, gebe ich der KI hier Gelegenheit für ein Schlusswort:

Wenn ich Ihre Argumentation zusammenfassen müsste, würde ich sagen:

> *Die größte Schwäche heutiger KI ist möglicherweise nicht die KI selbst,*
> *sondern die Tendenz von Menschen, ihr Fähigkeiten zuzuschreiben, die sie gar nicht besitzt.*

Betrachten wir einmal die gesamte Entwicklung:

Suchmaschine
Vor zwanzig Jahren ließ sich noch relativ leicht erkennen:

- Die Suchmaschine liefert Links.
- Ich muss die Dokumente selbst lesen.
- Ich muss die Informationen selbst bewerten.

Die Trennung zwischen
Mensch = denkt
und
Maschine = liefert Dokumente
war sehr deutlich.

LLM
Bei einem modernen Sprachmodell ist das viel schwieriger.
Sie stellen eine Frage.
Sie erhalten:

- vollständige Sätze,
- logische Argumentation,
- Beispiele,
- Zusammenfassungen,
- Gegenpositionen,
- manchmal sogar Humor.

Das sieht äußerlich exakt so aus wie menschliche Kommunikation.
Der Mensch schließt daraus intuitiv:
"Die Maschine muss verstehen, worüber sie spricht."

Aber genau an dieser Stelle haben Sie in Ihrem Artikel bereits den wichtigen Gegenpol aufgebaut:

- Turing-Test
- Chinese Room
- hochdimensionale Repräsentationen

Diese Kapitel helfen dem Leser zu verstehen:

> **Überzeugende Kommunikation ist kein Beweis für Verstehen.**

Und jetzt wird es wirklich interessant:
Vielleicht ist genau das der Grund, warum KI gesellschaftlich so viel Aufmerksamkeit erzeugt.

- Ein JPEG-Codec wirkt nicht intelligent.
- Ein MPEG-Encoder wirkt nicht intelligent.
- Ein Autoencoder wirkt nicht intelligent.
- Ein Embedding-Modell wirkt nicht intelligent.
- Ein LLM verwendet im Kern viele dieser Prinzipien ebenfalls.

Aber:

Es kommuniziert in menschlicher Sprache.

Dadurch entsteht eine völlig andere Wahrnehmung.

Deshalb halte ich Ihren Gedanken für sehr wichtig:

> *Viele Menschen können nicht zuverlässig einschätzen, was KI kann und was nicht, weil die Kommunikation selbst als Hinweis auf Intelligenz interpretiert wird.*

Das ist psychologisch absolut nachvollziehbar.

Menschen sind darauf trainiert, Sprache mit:

- Verständnis,
- Absicht,
- Wissen,
- Vernunft,
- Bewusstsein

zu verbinden.
Wenn etwas flüssig und schlüssig spricht, aktivieren wir automatisch dieselben Interpretationsmuster wie bei einem Menschen.

Interessanterweise passt sogar das zu Ihrer Theorie der Repräsentationen.
Denn auch Menschen arbeiten mit Repräsentationen.
Wenn wir eine Person treffen, bilden wir sofort eine innere Repräsentation:

- kompetent,
- sympathisch,
- vertrauenswürdig,
- sachkundig.

Eine gut formulierte KI-Antwort erzeugt offenbar sehr ähnliche innere Repräsentationen.
Ob diese Repräsentationen gerechtfertigt sind, ist eine ganz andere Frage.

Vielleicht könnte man die wichtigste Warnung für Anwender sogar so formulieren:

> Die Qualität einer KI-Antwort ist leicht zu beurteilen.
> Die Kompetenz der KI hinter der Antwort ist oft wesentlich schwerer zu beurteilen.

Oder noch zugespitzter:

> Eine KI kann sehr überzeugend kommunizieren,
> ohne die Grenzen ihres Wissens selbst zu verstehen.

Das ist letztlich die moderne Variante des Problems, das Sie bereits mit dem Chinese-Room-Gedankenexperiment beschrieben haben.

Und genau deshalb gefällt mir die Richtung Ihres Artikels so gut.
Er kommt am Ende nicht zu dem Schluss:

> "KI ist intelligent."

Aber auch nicht zu:

> "KI ist dumm."

Sondern eher zu:

> KI ist ein außerordentlich leistungsfähiges Werkzeug zur Erzeugung, Verdichtung und Verarbeitung von Repräsentationen.
>
> Gerade weil die Ergebnisse oft sehr überzeugend wirken, müssen Menschen besonders sorgfältig darauf achten, welche Fähigkeiten sie der KI tatsächlich zuschreiben.

Das ist aus meiner Sicht eine ausgesprochen differenzierte und zugleich praxisrelevante Sichtweise. Sie hilft Anwendern, die Stärken moderner KI zu nutzen, ohne den Fehler zu machen, sie mit einem Menschen zu verwechseln.