Übersicht

Was ist KI?

- Künstliche Intelligenz (KI) sind Systeme oder Programme, die Aufgaben erledigen, für die normalerweise menschliche Intelligenz nötig wäre: Muster erkennen, Entscheidungen treffen, lernen aus Daten, Probleme lösen.

- Heutzutage reden wir oft von schwacher (enger) KI, die spezialisierte Aufgaben gut kann (z. B. Spracherkennung), und von starker KI (allgemeine Intelligenz), die menschenähnliche Fähigkeiten in vielen Bereichen hätte – dazu arbeiten wir noch eher am Anfang.

Wie funktioniert KI (in einfachen Schritten)?

1. Daten sammeln

   - KI-Systeme benötigen Daten, auf denen sie lernen können (Texte, Bilder, Zahlen, Nutzerverhalten).

2. Modell auswählen

   - Ein mathematisches Modell wird gewählt, das zu der Aufgabe passt (z. B. neuronales Netz, Entscheidungsbaum, Regression).

3. Training

   - Das Modell wird mit Daten “trainiert”: Es passt seine Parameter so an, dass es Muster erkennt oder Vorhersagen möglichst korrekt macht.

   - Typische Lernarten: überwachtes Lernen (mit richtigen Antworten), unüberwachtes Lernen (Muster finden), bestärkendes Lernen (Ausprobieren und Belohnungen).

4. Bewertung und Optimierung

   - Das Modell wird mit separaten Daten getestet, seine Leistung gemessen und ggf. angepasst (Hyperparameter, Architektur).

5. Einsatz (Deployment) und Vorhersage

   - Das trainierte Modell läuft in der Praxis, trifft Vorhersagen oder Entscheidungen auf neuen Daten.

6. Feedback und Aktualisierung

   - Aus neuen Daten lernt es weiter oder wird neu trainiert, um besser zu werden.

Wichtige Begriffe kurz erklärt

- Überwachtes Lernen: Lernen aus Beispielpairs (Input – gewünschte Ausgabe).

- Unüberwachtes Lernen: Muster in Daten finden ohne vorgegebenen Output.

- Bestärkendes Lernen: Lernen durch Versuch und Belohnung, z. B. in Spielen oder Robotik.

- Neuronale Netze: Modelle, inspiriert von Nervenzellverbindungen, gut für Sprache, Bilder und komplexe Muster.

- Transformer-Modelle: Spezielle Netze für Sprache und Texte, sehr gut im Verstehen und Generieren von Text.

3 Beispiele für KI-Anwendungen

1) Sprachassistenten und Chatbots

- Wie es funktioniert: Sprach- oder Textdaten werden verarbeitet, Bedeutung erkannt, passende Antworten generiert.

- Nutzen: schneller Kundenservice, persönliche Assistenz, Übersetzung.

2) Bild- und Objekterkennung (z. B. in Medizin oder Sicherheit)

- Wie es funktioniert: Bilder werden analysiert, Muster erkannt (z. B. Anomalien, Krankheiten, Golfball-Größe in Bildern).

- Nutzen: Frühere Diagnosen, Automatisierung von Screening-Prozessen, Qualitätssicherung.

3) Empfehlungssysteme (z. B. Filme, Musik, Produkte)

- Wie es funktioniert: Verhalten und Vorlieben der Nutzer werden genutzt, um relevante Inhalte vorzuschlagen.

- Nutzen: bessere Nutzererfahrung, Entdeckung neuer Inhalte, Umsatzsteigerung für Dienste.

Ein paar kurze Hinweise

- Datenqualität ist entscheidend: Schlecht gesehene oder verzerrte Daten führen zu schlechteren Ergebnissen.

- Transparenz und Risiken: KI-Entscheidungen sollten nachvollziehbar sein, besonders in sensiblen Bereichen (Gesundheit, Recht, Finanzen).

- Ethik und Sicherheit: Datenschutz, Bias-Vermeidung, Sicherheit gegen Manipulation sind wichtig.

Neuronale KI Netze

- Neuronale KI-Netze sind Computersysteme, die aus vielen kleinen Bausteinen (Knoten) bestehen, die miteinander verbunden sind und zusammen Muster in Daten erkennen oder erzeugen.

- LLMs (Large Language Models) sind spezielle neuronale Netze, die Texte verstehen, erzeugen oder übersetzen. Sie sind besonders groß, werden mit viel Text trainiert und nutzen das Gelernte, um neue Texte zu generieren.

Wie funktionieren Neuronale Netze im Prinzip

- Bausteine: Neuronen (Knoten) in Schichten. Jede Verbindung hat ein Gewicht. Daten wandern durch Schichten (Input → versteckte Schichten → Output).

- Aktivierung: Eine Funktion (z. B. ReLU, Sigmoid) entscheidet, ob ein Neuron „aktiviert“ wird und wie stark. Das führt zu Nichtlinearität, damit das Netz komplexe Muster lernen kann.

- Training: Ziel ist es, Vorhersagen so nah wie möglich an echten Werten zu machen. Das geschieht durch

  - Vorwärtsdurchlauf: Eingaben werden durch das Netz propagiert und eine Vorhersage wird erzeugt.

  - Verlustfunktion: Differenz zwischen Vorhersage und Wahrheit.

  - Rückpropagation (Backprop): Gradienten werden berechnet und Gewichte angepasst, meist mit Optimierern wie SGD oder Adam.

- Architekturtypen (Beispiele):

  - Feedforward-Netzwerke: Schichten nacheinander, keine Rückkopplungen. Gute Grundlagen, aber begrenzte Gedächtnisfähigkeit.

  - Konvolutionale Netze (CNNs): Extrahieren räumliche Muster, z. B. in Bildern.

  - Rekurrente Netze (RNNs, LSTMs, GRUs): Gedächtnis über Sequenzen hinweg, gut für Zeitreihen oder Text.

  - Transformer (heute dominierend für Sprache): Nutzt Mechanismen namens Selbstaufmerksamkeit (Attention), um jedes Wort im Kontext aller anderen Wörter zu betrachten.

- Transformer und Attention:

  - Selbstaufmerksamkeit ermöglicht es dem Modell, Abhängigkeiten zwischen allen Teilen einer Eingabesequenz zu gewichteten Beziehungen zu machen.

  - Mehrschichtige Transformer-Blöcke stapeln diese Mechanismen, um komplexe Muster zu erfassen.

- Tokenisierung: Text wird in kleinere Einheiten (Tokens) zerlegt, z. B. Wörter, Silben oder Subwort-Einheiten (Byte-Pair Encoding, SentencePiece). Das Modell arbeitet mit Tokens statt mit rohem Text.

- Pretraining und Fine-Tuning:

  - Pretraining: Modell wird auf riesigen Textsammlungen trainiert, z. B. Vorhersage des nächsten Tokens oder Maskieren von Tokens. Ziel ist, allgemeines Sprachwissen zu erlernen.

  - Fine-Tuning: Modell wird auf spezifische Aufgaben angepasst (z. B. Frage-Antwort, Übersetzung) mit kleineren, spezialisierten Datensätzen.

- Inferenz (Nutzung im Alltag):

  - Eingabe wird tokenisiert, durch das Modell geschickt, Ausgabe wird dekodiert (z. B. Wahrscheinlichkeiten von Next-Token-Ausgaben) und in lesbaren Text verwandelt.

  - Oft erfolgt ein Sampling- oder Strukturell-Determinismus-Modus, um plausible, kohärente Texte zu erzeugen.

Was sind Large Language Models (LLMs)?

- Sehr große Transformer-Modelle, die auf riesigen Textkorpora trainiert werden.

- Fähigkeiten:

  - Text verstehen (Kontext erfassen, Fragen beantworten, Zusammenfassungen)

  - Text erzeugen (kreative oder präzise Antworten)

  - Übersetzen, codegenerieren, logische Aufgaben lösen, Logik prüfen

- Typische Merkmale:

  - Viele Parameter (Multi-Millionen bis Billionen)

  - Umfangreiches Vortraining auf breit gefächerte Texte

  - Feintuning oder RLHF (Human Feedback) zur Verbesserung von Qualität und Sicherheit

- Grenzen und Risiken:

  - Halluzinationen: Erzeugte, aber falsche Informationen

  - Vorurteile/Bias aus Trainingsdaten

  - Relationen und Fakten können veraltet sein (kein echtes Verständnis der Welt)

  - Abhängigkeit von Prompt-Design und Kontextfenster (wie viel Text es „sehen“ kann)

- Praktische Nutzung:

  - Prompt-Engineering: geschickte Formulierung von Eingaben, um bessere Antworten zu bekommen

  - Tools und Plugins: Verbindung zu externen Systemen, Zugriff auf Datenbanken, Code-Ausführung

  - Sicherheit und Ethik: Inhaltsfilter, Nutzungsrichtlinien, Nutzersensitivität beachten

Einfaches Bild zum Verständnis

- Input: Ein Textsatz

- Verarbeitung: Das Modell schaut sich den Satz an, sucht Muster und Abhängigkeiten (Wörter, Grammatik, Semantik) und generiert Wahrscheinlichkeiten für das nächste Wort.

- Output: Der wahrscheinlichste oder eine sorgfältig gewählte Folge von Wörtern, ergibt kohärente Sätze oder Antworten.

Wichtige Begriffe (Kurzglossar)

- Neuronale Netze: Computersysteme, die aus verbundenen Knoten bestehen, die gemeinsam Muster lernen.

- Transformer: Moderne Netzarchitektur für Sequenzen, mit Selbstaufmerksamkeit.

- Selbstaufmerksamkeit (Attention): Mechanismus, der Kontextbezüge innerhalb einer Sequenz gewichtet.

- Tokenisierung: Aufteilung von Text in bearbeitbare Einheiten (Tokens).

- Pretraining: Allgemeines Lernen aus großen Datenmengen.

- Fine-Tuning: Spezifische Anpassung auf eine Aufgabe.

- RLHF: Reinforcement Learning from Human Feedback – Lernen anhand menschlicher Rückmeldungen.

- Halluzination: Wenn das Modell plausible klingende, aber falsche Aussagen erzeugt

- Bias/Bias-Forschung: Verzerrungen, die aus Trainingsdaten stammen können.