1 Introduction to AI Principle

1.1 Was ist Künstliche Intelligenz?

Die Definition von Künstlicher Intelligenz (KI) ist nicht eindeutig und unterliegt einem stetigen Wandel („Moving Goalposts“). Aufgaben, die früher als KI galten, werden heute oft als reine Softwaretechnik angesehen, sobald sie gelöst sind.

Definitionen von KI

Es gibt verschiedene Ansätze, KI zu definieren:

- **John McCarthy (1971):** „Die Wissenschaft und Ingenieurskunst, **intelligente Maschinen** zu bauen, insbesondere intelligente Computerprogramme.“ Er betonte, dass KI nicht auf biologisch beobachtbare Methoden beschränkt sein muss.
- **Marvin Minsky (1969):** „Die Wissenschaft, Maschinen dazu zu bringen, Dinge zu tun, die **Intelligenz** erfordern würden, wenn sie von Menschen getan würden.“

Gründe für die schwierige Definition sind unter anderem:

- Fehlende offizielle Definition.
- Einfluss von Science-Fiction auf die öffentliche Wahrnehmung.
- Das Phänomen, dass ehemals schwere Aufgaben (z.B. Schach) nach ihrer Lösung als „einfache Rechnerei“ abgetan werden, während Aufgaben, die für Menschen einfach sind (z.B. Wahrnehmung, Motorik), für Maschinen sehr schwer sein können (Moravec's Paradox).

1.2 Intelligenztests und Philosophie

Um festzustellen, ob ein System intelligent ist, wurden verschiedene Tests und Gedankenexperimente entwickelt.

1.2.1 Der Turing-Test (Imitation Game)

Vorgeschlagen von Alan Turing. Die Grundannahme ist, dass ein Wesen intelligent ist, wenn es sich in seinem Verhalten nicht von einem anderen intelligenten Wesen unterscheiden lässt.

Ablauf:

1. Ein menschlicher Fragesteller interagiert blind (per Text) mit zwei Spielern: A und B.
2. Einer der Spieler ist ein Mensch, der andere ein Computer.
3. Wenn der Fragesteller nicht zuverlässig entscheiden kann, welcher Spieler der Computer ist, hat der Computer den Test bestanden.

Kritik: Der Test prüft *Verhalten*, nicht das *Verständnis* oder *Bewusstsein*. Ein System kann menschliches Verhalten simulieren, ohne die zugrunde liegenden Konzepte zu verstehen.

1.2.2 Das Chinesische Zimmer (The Chinese Room)

Ein Gegenargument zum Turing-Test von John Searle, das den Unterschied zwischen Syntax (Zeichenmanipulation) und Semantik (Bedeutung) hervorhebt.

Szenario:

- Eine Person, die kein Chinesisch versteht, sitzt in einem geschlossenen Raum.
- Sie hat ein Regelbuch (Programm), das beschreibt, wie auf chinesische Zeichenfolgen (Input) mit anderen chinesischen Zeichenfolgen (Output) reagiert werden soll.

- Für einen Außenstehenden wirkt es so, als verstünde die Person Chinesisch.

Schlussfolgerung: Die Person manipuliert nur Symbole anhand ihrer Form (Syntax), versteht aber deren Inhalt (Semantik) nicht. Ebenso könnte eine KI intelligent *wirken* (Turing-Test bestehen), ohne wirklich intelligent zu *sein*.

1.3 Kategorisierung von KI

1.3.1 Starke vs. Schwache KI

Unterscheidung der Reichweite

- **Generelle KI (General AI / Strong AI):** Ein System, das **jede** intellektuelle Aufgabe bewältigen kann, die auch ein Mensch lösen kann. Es besitzt Verständnis und Bewusstsein (Forschungsziel).
- **Schwache KI (Narrow AI / Weak AI):** Ein System, das darauf spezialisiert ist, eine **konkrete** oder eine begrenzte Menge von Aufgaben zu lösen (aktueller Stand der Technik, z.B. Schachcomputer, Bilderkennung).

1.3.2 Eigenschaften eines KI-Systems

Ein modernes KI-System sollte idealerweise folgende Eigenschaften aufweisen:

- **Adaptability (Anpassungsfähigkeit):** Die Fähigkeit, die Leistung durch Lernen aus Erfahrung zu verbessern.
- **Autonomy (Autonomie):** Die Fähigkeit, Aufgaben in Umgebungen ohne ständige Anleitung durch einen Benutzer oder Experten auszuführen.
- **Rationality (Rationalität):** Das Treffen der „richtigen“ Entscheidungen (siehe unten).

1.4 Dimensionen der KI-Definition

KI kann entlang zweier Dimensionen klassifiziert werden:

1. **Prozess:** Fokus auf Denkprozesse/Schlussfolgern vs. Fokus auf Verhalten/Handeln.
2. **Maßstab:** Erfolg gemessen am menschlichen Vorbild vs. Erfolg gemessen an einem idealen Rationalitätsbegriff.

Dies ergibt vier Felder der KI-Forschung:

	Menschlicher Maßstab	**Ideal (Rationalität)**
Denken	**Systems that think like humans** (Kognitionswissenschaft: Modellierung der menschlichen Denkweise)	**Systems that think rationally** (Logik: „Gesetze des Denkens“, korrekte Schlussfolgerungen)
Handeln	**Systems that act like humans** (Turing-Test Ansatz: Simulation menschlichen Verhaltens)	**Systems that act rationally** (Rationaler Agent: Maximierung des erwarteten Nutzens)

Table 1: *Die vier Ansätze der KI*

1.4.1 Rationalität vs. Gesetze des Denkens

- **Laws of Thought (Logik):** Beschäftigt sich mit unwiderlegbaren, logischen Schlussfolgerungen. Problem: In der Realität gibt es oft Unsicherheiten, für die Logik allein nicht ausreicht.
- **Rational Behavior (Rationales Handeln):** Das „Richtige“ tun. Das Richtige ist definiert als das, was die Erreichung der Ziele angesichts der verfügbaren Informationen maximiert (Maximierung des *Expected Utility*).
- **Vorteile der Rationalität:** Sie ist allgemeiner als reine Logik (funktioniert auch bei Unsicherheit) und wissenschaftlich besser handhabbar (Optimierungsproblem).

1.5 Grundlagen und verwandte Disziplinen

KI ist ein interdisziplinäres Feld, das auf vielen Bereichen aufbaut:

- **Philosophie:** Logik, Reasoning, Geist als physisches System, Grundlagen des Lernens.
- **Mathematik:** Formale Repräsentation, Beweise, Algorithmen, Wahrscheinlichkeitstheorie.
- **Psychologie / Kognitionswissenschaft:** Wahrnehmung, Motorik, Anpassung. Kognitionswissenschaft verbindet KI-Modelle mit experimentellen Techniken der Psychologie.
- **Ökonomie:** Entscheidungstheorie, Spieltheorie (rationale Entscheidungen).
- **Neurowissenschaften:** Physisches Substrat (Gehirn) für mentale Aktivitäten.
- **Linguistik:** Wissensrepräsentation, Grammatik.
- **Kontrolltheorie:** Stabilität, optimales Agentendesign.

1.5.1 Taxonomie der KI

Die Begriffe werden oft hierarchisch verstanden:

- **Künstliche Intelligenz (KI):** Der Überbegriff für Technik, die intelligente Züge zeigt.
- **Machine Learning (Maschinelles Lernen):** Ein Teilgebiet der KI, das Algorithmen nutzt, um aus Daten zu lernen (statt explizit programmiert zu werden).
- **Deep Learning:** Ein Teilgebiet des Machine Learning, das auf künstlichen neuronalen Netzen mit vielen Schichten basiert (aktuell sehr erfolgreich, siehe Nobelpreis Physik 2024 an Hinton/Hopfield).

1.6 Grenzen und Probleme moderner KI

Trotz großer Erfolge (z.B. AlphaGo, Stable Diffusion, Protein Folding) existieren signifikante Limitationen:

- **Bias (Voreingenommenheit):** KI-Modelle übernehmen Vorurteile aus den Trainingsdaten (z.B. rassistische Tendenzen in Gesichtserkennung oder Textgenerierung).
- **Adversarial Attacks:** KI kann leicht getäuscht werden. Durch für Menschen unsichtbare Änderungen an einem Bild (Rauschen) kann eine KI dazu gebracht werden, ein Objekt völlig falsch zu klassifizieren (z.B. Panda wird mit 99% Konfidenz als Gibbon erkannt).
- **Halluzinationen:** Generative KIs erzeugen plausibel klingende, aber faktisch falsche Informationen.
- **Isolation:** Aktuelle KI ist meist „Narrow AI“ und auf spezifische Probleme beschränkt, ohne allgemeines Weltverständnis.

1.7 Ergänzung: Agenten (aus der Übung)

Ein zentrales Konzept der KI ist der **Agent**.

- **Agent:** Eine Einheit, die ihre Umgebung über **Sensoren** wahrnimmt und mittels **Aktuatoren** auf diese Umgebung einwirkt.
- **Environment (Umgebung):** Die Welt, in der der Agent operiert.

Zur Beschreibung eines KI-Problems (z.B. Roboterfußball) werden oft folgende Aspekte analysiert:

- **Performance Measure:** Wie wird Erfolg gemessen? (z.B. Anzahl der Tore, gewonnene Spiele).
- **Environment:** Was beinhaltet die Welt? (z.B. Spielfeld, Ball, Gegner, Tore).
- **Actuators:** Wie kann der Agent handeln? (z.B. Motoren für Beine, Schussmechanik).
- **Sensors:** Wie nimmt der Agent wahr? (z.B. Kameras, Beschleunigungssensoren).

Umgebungseigenschaften können klassifiziert werden als:

- *Observable* (vollständig beobachtbar) vs. *Partially Observable*.

- *Accessible* (Kann ein Agent vollständige & akkurate Informationen erhalten (gibt es diese überhaupt)) vs. *Inaccesssible*
- *Deterministic* (Folgezustand durch aktuellen Zustand und Aktion bestimmt) vs. *Stochastic*.
- *Episodic* (Handlungen sind unabhängig voneinander) vs. *Sequential*.
- *Static* (Umgebung ändert sich nicht während der Entscheidungsfindung) vs. *Dynamic*.
- *Discrete* (endliche Anzahl an Zuständen/Aktionen) vs. *Continuous*.