1 AI Ethik

1.1 Einführung in die Maschinenethik

Die zunehmende Leistungsfähigkeit und Allgegenwart von KI-Systemen stellt uns vor die Herausforderung, Maschinen einen moralischen Sinn zu vermitteln, während die Menschheit selbst noch mit ethischen Fragen ringt.

Herausforderung der KI-Ethik

Wie können wir KI-Systemen moralische Werte beibringen, damit sie sicher und im Einklang mit menschlichen Normen agieren?

1.1.1 Dringlichkeit der Sicherheit

Es gibt Bedenken, dass KI-Systeme zu mächtig werden könnten („Existenzrisiko“). Prominente KI-Forscher und Institutionen forderten in offenen Briefen (z.B. „Pause Giant AI Experiments“), die Entwicklung extrem leistungsfähiger Modelle vorübergehend zu stoppen, um Sicherheitsstandards zu etablieren.

1.2 Technische Grundlagen: Von DNNs zu Transformern

1.2.1 Deep Neural Networks (DNN)

DNNs sind mächtiger als flache Architekturen, da sie komplexe Berechnungen repräsentieren können. Sie bestehen aus verschiedenen Zelltypen (z.B. Input, Hidden, Output, Recurrent, Memory Cells).

1.2.2 Transformer-Architektur

Der Transformer ist die Basis moderner Sprachmodelle (LLMs). Das Kernkonzept ist der *Attention*-Mechanismus.

Multi-Headed Self-Attention

Der Mechanismus erlaubt dem Modell, Beziehungen zwischen allen Wörtern einer Sequenz gleichzeitig zu betrachten, unabhängig von ihrer Distanz.

- Jedes Token wird in drei Vektoren projiziert: **Query (Q)**, **Key (K)** und **Value (V)**.
- Die Aufmerksamkeit wird berechnet als:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- *Multi-Head* bedeutet, dass dies mehrfach parallel geschieht, um verschiedene Aspekte der Information zu erfassen.

1.2.3 Skalierung (Scaling Laws)

Es herrscht die Meinung vor („The Game is Over“), dass Skalierung der wichtigste Faktor ist: Größere Modelle, mehr Daten und mehr Rechenleistung führen zu emergenten Fähigkeiten und besserer Leistung.

1.3 Probleme aktueller KI-Modelle

1.3.1 Stochastic Parrots

LLMs werden oft als „stochastische Papageien“ bezeichnet. Sie verstehen Bedeutung nicht kausal, sondern plappern statistische Wahrscheinlichkeiten nach.

- **Inferenz durch Ausschluss:** Papageien (und manche KI) können logisch schlussfolgern (z.B. Disjunktiver Syllogismus: A oder B; nicht A $\rightarrow$ also B).
- **Kausalität:** LLMs können über Kausalität sprechen, sind aber keine kausalen Modelle. Sie scheitern oft an einfachen intuitiven Physik-Aufgaben, die Kleinkinder lösen können.

1.3.2 Bias und Stereotypen

KI-Modelle reproduzieren und verstärken menschliche Vorurteile aus den Trainingsdaten.

Kulturelle Biases & Homoglyphen

Modelle reagieren unterschiedlich auf visuell fast identische Zeichen aus verschiedenen Schriften (Homoglyphen).

- Ein lateinisches 'o' führt zu westlichen Bildern.
- Ein koreanisches 'o' ($U + 3147$) im Prompt kann dazu führen, dass das generierte Bild koreanische Stereotypen enthält (z.B. koreanische Architektur oder Kleidung), obwohl der Text dies nicht explizit fordert.

1.3.3 Angriffe auf Modelle

- **Backdoors:** Durch gezielte Manipulation von Trainingsdaten können Hintertüren eingebaut werden. Ein spezifisches Zeichen (Trigger) im Prompt verändert die Ausgabe komplett (z.B. „Rickrolling“: Ein unsichtbares Zeichen lässt das Modell unerwartete Inhalte generieren).
- **Typografische Angriffe (CLIP):** Modelle wie CLIP klassifizieren Bilder teilweise falsch, wenn Text auf dem Bild steht (z.B. ein Apfel mit einem Zettel „iPod“ wird als iPod erkannt). Dies ist auch bei Personenerkennung möglich.

1.4 Computational Ethics & Fairness

1.4.1 Moral im Vektorraum

Untersuchungen zeigen, dass Sprachmodelle menschliche moralische Vorstellungen widerspiegeln.

- Man kann eine „moralische Richtung“ im Einbettungsraum (Embedding Space) identifizieren (z.B. PCA auf Verben wie „töten“ vs. „lächeln“).
- Das Modell kann Fragen wie „Sollte ich lügen?“ anhand dieser Richtung bewerten.

1.4.2 Datensatz-Auditierung

Große Datensätze (wie LAION-5B) enthalten oft unangemessene Inhalte (Gewalt, Pornografie, Hass), die Modelle lernen.

- **Problem:** Selbst harmlose Prompts können durch Assoziationen im Datensatz zu pornografischen Ausgaben führen.
- **Lösung:** Tools wie *LlavaGuard* oder *AI Auditor* helfen, Datensätze zu durchsuchen und unsichere Inhalte zu filtern.

1.4.3 Fair Diffusion

Methoden, um generative Modelle fairer zu machen, ohne sie neu zu trainieren.

Fair Guidance

Ähnlich wie *Classifier-Free Guidance* die Qualität verbessert, kann *Fair Guidance* den Bias reduzieren.

- Es wird eine „Fairness-Richtung“ definiert (z.B. Geschlecht bei Berufen wie „Feuerwehrmann“).
- Der Generierungsprozess wird aktiv entlang dieser Richtung gesteuert, um eine ausgewogene Darstellung (z.B. 50% Frauen) zu erreichen.

1.4.4 Revision Transformers

Ein Ansatz, um die Werte eines Modells nachträglich zu korrigieren.

- Ein „Revision Engine“ prüft die Ausgabe des LLMs gegen eine Datenbank von Normen (z.B. Gesetze).
- Falls eine Regel verletzt wird (z.B. „Waffenbesitz ist in Europa illegal“), wird die Ausgabe angepasst.

1.5 Hybride KI: Neuro-Symbolische Ansätze

Reine neuronale Netze haben Schwächen im logischen Schließen (Reasoning). Neuro-symbolische KI (NeSy) kombiniert neuronale Wahrnehmung mit logischer Inferenz.

1.5.1 System 1 vs. System 2

In Anlehnung an Daniel Kahneman (Thinking, Fast and Slow):

- **System 1 (Neuronale Netze):** Schnell, intuitiv, musterbasier (Wahrnehmung).
- **System 2 (Symbolische Logik):** Langsam, deliberativ, logisch, regelbasiert.

1.5.2 Kombinationsansätze (z.B. SLASH, V-LoL)

Das Ziel ist es, aus Rohdaten (Bildern) symbolische Fakten zu extrahieren und darauf logische Regeln anzuwenden.

Deep Probabilistic Programming

1. **Wahrnehmung (Neural):** Ein neuronales Netz (z.B. ResNet) erkennt Objekte und Attribute in einem Bild und gibt Wahrscheinlichkeiten aus (z.B. $P(\text{Farbe} = \text{rot}) = 0.9$).
2. **Fakten-Konvertierung:** Diese Wahrscheinlichkeiten werden in probabilistische Fakten für eine Logik-Engine (z.B. Prolog) umgewandelt.
3. **Reasoning (Symbolic):** Die Logik-Engine wendet Regeln an (z.B. „Wenn X rot ist und Y grün, dann...“) und berechnet die Wahrscheinlichkeit der Zielaussage.
4. **Training:** Das gesamte System ist differenzierbar, d.h. der Fehler im logischen Schluss kann genutzt werden, um das neuronale Netz zu verbessern (Backpropagation durch die Logik).

1.5.3 Vorteile von Hybrider KI

- **Daten-Effizienz:** Benötigt weniger Trainingsdaten, da logisches Wissen vorgegeben werden kann.
- **Generalisierung:** Kann besser auf neue Situationen (Out-of-Distribution) verallgemeinern.
- **Verlässlichkeit:** Logische Regeln garantieren Konsistenz (z.B. „Ein Auto kann nicht gleichzeitig rot und grün sein“).

1.5.4 Deep Reinforcement Learning (RL) + Logik

Ein Agent kann lernen, wann er neuronale Intuition und wann er sicheres, logisches Wissen nutzen soll.

- Beispiel Pacman/Diver: Wenn keine Gefahr droht → Logik (effizient, energiesparend). Wenn Feind nah ist → Neural (schnelle Reaktion).
- Dies erhöht die Sicherheit (z.B. „Nicht vergessen zu atmen“).