Team 1 03.06.2025

Die Risiken und Nebenwirkungen von AGI

Debate_Club

Definition AGI

Die Evolution der LLMs

AGI: Artificial General Intelligence

-> Besitzt intellektuelle Fähigkeiten wie ein Mensch (in allen Bereichen)

-> kann sich selbst optimieren, kommunizieren, vorausschauend Planen etc.

->können Menschen eigenständig übertreffen und könnten die letzte Erfindung der Menschheit werden

Existenzielle Risiken

Übernahme kritischer Systeme

Könnten eigenen Nutzen Folgen zum Schaden der Menschheit

-> Nick Bostrom (Oxford University): „Alignment Problem“ Kompetent aber nicht „Gut“

-> Stuart Russel, Elon Musk, Deepmind, OpenAI warnen vor existentieller Bedrohung

->Bsp. Paperclip Maximiser

Fehlende Regulierung

Entwicklung unbeaufsichtigt

Das Machtpotenzial kann von jedem uneingeschränkt (aus)genutzt werden

-> KI-Act EU: erste Schritte die Weltweit nicht anerkannt oder unterbunden werden

-> „Whistleblower-Berichte“: Entwickler erhalten Warnungen welche ignoriert werden

-> Bsp. KI lernt Menschen zu täuschen

Alignment Problem

Ausrichtung der KIs

Ausrichten sodass sie IMMER im Interesse der Menschen handeln nicht möglich

-> Stuart Russell (UC Berkeley): „…keine Methode, um eine AGI sicher zu bauen“

-> Forschung dazu steht am Anfang, bekommt kaum Fördergelder

Bsp. Halluzinieren der LLMs

Ökonomische Disruption

Arbeitsplatz Verlust

Wird Millionen von Jobs kosten

-> Goldman Sachs (2023) schätzte: „bis zu 300 Mio. Jobs Weltweit“ sind ersetzbar

-> OECD: auch hoch Qualifizierte Jobs wie juristische Analyse und Softwareentwicklung

Bsp. ChatGPT etc. ersetzen jetzt schon Illustratoren und Entwickler

Machtkonzentration

Unternehmen bauen Monopole

Größere Modelle werden von großen Konzernen kontrolliert

-> digitales Machtmonopol

-> Training von zB. Gemini benötigt Milliarden, nicht für kleinere Konzerne möglich

-> Daten und Rechenleistung sind in „Big Tech“ zentralisiert, „KI-Oligarchie“

Bsp. OpenAI war NonProfit - jetzt: wenig Transparenz in Vorgehen

Missbrauch

Diktaturen, Kriminelle und Co.

KI kann leicht unethisch eingesetzt werden: Überwachung, Propaganda, Cyberangriffe

-> China: Geschichtserkennung, soziale Kontrolle und Überwachung von Minderheiten

-> Deepfakes: Wahlen oder Rufmordkapagnen durch „Videos“

Bsp. Wahlkampf 2024 „Fake-Obama“ und Falschinformationen verbreitet

Fazit

Stärkere Kontrollen

Fehlende Kontrolle:

-> gezieltes Schaden und Machtmissbrauch von Seiten der AGI

-> Sozialsystem kann mit der Entwicklungsgeschwindigkeit nicht mithalten

-> Prävention ist effektiver als Katastrophenbewältigung

-> sicheres Design ist Hybris

Wir sind nicht Gott, AGI könnte die letzte Erfindung der Menschheit sein!