

---

# **Philosophie der AI**

Gerd Graßhoff

2024-06-01

# Table of contents

<b>Philosophie der AI</b>	<b>8</b>
<b>1 Was ist AI?</b>	<b>9</b>
1.1 Begrüßung und Einführung . . . . .	9
1.2 AI als Alleskönner . . . . .	11
1.2.1 Der Durchbruch der AI-Visionen . . . . .	11
1.2.2 Die Attraktivität von AI . . . . .	12
1.2.3 Die ursprüngliche Idee des Internets . . . . .	12
1.2.4 Die Ablösung der Webwelt durch AI . . . . .	12
1.2.5 Die Umgestaltung der Architektur des Internets . . . . .	13
1.3 Neue Möglichkeiten durch Künstliche Intelligenz . . . . .	14
1.3.1 Hochwertige Übersetzungen . . . . .	14
1.3.2 Simultanübersetzung und Lektoratsassistenz . . . . .	14
1.3.3 Automatisierte Forschungsberichte . . . . .	15
1.3.4 Das Labor <i>Lettre AI</i> . . . . .	15
1.3.5 Übertragen eines Bildes in maschinenlesbaren Text . . . . .	17
1.3.6 Übersetzen des Textes in eine andere Sprache . . . . .	17
1.4 Erweiterungen . . . . .	17
1.4.1 Analogie zu Sherlock Holmes . . . . .	18
1.4.2 Vielfältige Analysemöglichkeiten von Texten . . . . .	18
1.5 Philosophie als Grundlage für die Möglichkeiten der AI . . . . .	18
1.5.1 Beantwortung von Fragen über Mikrofoneingabe . . . . .	18
1.5.2 Die Möglichkeiten der AI . . . . .	19
1.5.3 Gefahren der AI . . . . .	19
1.5.4 Der sprachliche Kern der AI . . . . .	20
1.5.5 Das Problem der Halluzinationen . . . . .	20
1.5.6 Die Gefahr der Manipulation durch glaubwürdige Fakes . . . . .	21
1.5.7 Selektive Informationen und die Pluralität der Hintergründe . . . . .	21

1.5.8	Die Unausweichlichkeit der AI-Entwicklung und die Notwendigkeit der Gestaltung	21
1.5.9	Weitere Gefahren: Diskriminierung und Überwachung	22
1.5.10	Die Notwendigkeit der Auseinandersetzung mit AI	22
1.6	Nutzungsmöglichkeiten in der Wissenschaft	22
1.7	Bislang nicht lösbarer Aufgaben	23
1.7.1	Frage 1: Einfache Aussage in einer Quelle	23
1.7.2	Frage 2: Aussage in Briefen zu einem Thema	24
1.7.3	Frage 3: Aussagen einer Person in ihren Schriften	24
1.7.4	Frage 4: Keine Aussage einer Person in ihren Schriften	24
1.8	Die Herausforderung der inhaltlichen Analyse mit AI	25
1.8.1	Grenzen der traditionellen Datenbanken	26
1.8.2	Qualifizierte Aussagen auf Basis der verfügbaren Evidenz	26
1.8.3	Herausforderungen bei der Interpretation von Metaphern und Ironie	26
1.8.4	Lernfähigkeit und Entwicklungspotenzial von AI-Systemen	27
1.8.5	Der Paradigmenwechsel durch Large Language Models und Embeddings	27
1.8.6	Die Bedeutung der Philosophie für die AI-Forschung	28
<b>2</b>	<b>Die Revolution der AI</b>	<b>29</b>
2.1	Begrüßung und Rückblick auf die letzte Vorlesung	29
2.2	Traditionell schwer lösbarer Fragen in der Forschung	29
2.2.1	Noch schwieriger: Evidenz zur Widerlegung von Hypothesen finden	29
2.2.2	Komplexe Fragen zur zeitgenössischen Rezeption historischer Hypothesen	30
2.3	Die Bedeutung der AI für die Geisteswissenschaften	30
2.4	Die Evolution der Mensch-Maschine-Interaktion	30
2.4.1	Von der Adresseingabe zur Suchanfrage	30
2.4.2	Der Durchbruch von Chat-GPT	31
2.4.3	Neue Schnittstellen: Sprache, Gesten und Gedanken	31
2.5	Die Macht der generativen AI	31
2.5.1	Von der Syntax zur Semantik	31
2.5.2	Die Bedeutung sprachlicher Ausdrücke	31
2.5.3	Philosophische Kritik an der Terminologie	32
2.5.4	Von der Zeichenkettensuche zur Bedeutungsanalyse	32
2.5.5	Wahrheitswerte und die Welt der Aussagen	32
2.5.6	Die Dimension der Aussagen eröffnet neue Möglichkeiten	33

2.6	Die drei Säulen der semantischen Revolution . . . . .	33
2.6.1	1. Das Training mit bedeutungsähnlichen Begriffen . . . . .	34
2.6.2	2. Die Frage nach der Bedeutungsgleichheit . . . . .	34
2.6.3	3. Das Training mit logischen Regeln . . . . .	35
2.7	Ausblick . . . . .	36
2.7.1	Die Bedeutung von Embeddings . . . . .	36
2.7.2	Die Suche nach bedeutungsähnlichen Aussagen . . . . .	36
2.7.3	Die Erweiterung auf verschiedene Medien . . . . .	37
2.8	Die zweite Revolution: Attention is all you need . . . . .	37
2.8.1	Die Macht der Vorhersage . . . . .	37
2.8.2	Von der Frage zur Anweisung . . . . .	37
2.9	Die Komposition von Instruktionen und Inhalten . . . . .	38
2.9.1	Sprachkompetenz vs. Sachkompetenz . . . . .	38
2.9.2	Gefahren und Grenzen von Chat-GPT . . . . .	39
2.10	Erweiterung der AI-Modelle . . . . .	39
2.10.1	Sachliche Korrektheit und Wahrheit . . . . .	39
2.10.2	Korrespondenztheorie der Wahrheit . . . . .	40
2.11	Sprachentwicklung und Bedeutungsverschiebungen . . . . .	40
2.11.1	Fehltraining und Sprachmarotten . . . . .	40
2.11.2	Reichhaltige Kontextkonstruktion . . . . .	41
2.12	Anwendungsbeispiele und Potenziale . . . . .	41
2.12.1	Übersetzungen als Motor des Trainings . . . . .	41
2.12.2	Zusammenfassungen und Frage-Antwort-Systeme . . . . .	42
<b>3</b>	<b>Charakter von LLMs</b>	<b>43</b>
3.1	Vorlesung Philosophie der AI: Generative Modelle, Large Language Models und Character-Konfiguration . . . . .	43
3.2	Die Revolution der generativen AI-Modelle . . . . .	43
3.2.1	Large Language Models als Kern der generativen AI . . . . .	44
3.2.2	Die Explosion der verfügbaren Modelle . . . . .	44
3.3	Die Funktionsweise der generativen AI-Modelle . . . . .	44
3.3.1	Semantische Ähnlichkeit und Transformation . . . . .	44
3.3.2	Character - Die Formung des künstlichen Charakters . . . . .	44
3.3.3	Metaregeln und kausales Schließen . . . . .	45
3.3.4	Historisches Schließen . . . . .	46

3.3.5	Die Bedeutung des Kontexts . . . . .	46
3.3.6	AGI - Ein umstrittenes Konzept . . . . .	47
3.3.7	Hermeneutik als Herausforderung für AI . . . . .	48
3.3.8	Kontextvergrößerung als Schlüssel zum Verständnis . . . . .	48
3.3.9	Ausblick . . . . .	49
3.3.10	Von der Query zur Instruktion . . . . .	50
3.4	Die Schlüsselemente der Revolution: Semantische Ähnlichkeit und regelhafte Textgenerierung . . . . .	50
3.4.1	Wer war Johann Wolfgang Goethe? - Eine typische Google-Frage . . . . .	50
3.4.2	Die Grenzen der Aktualität . . . . .	51
3.4.3	Interne Präferenzordnungen und Regeln . . . . .	51
3.5	Die Qualität der Internetressourcen reicht nicht aus . . . . .	51
3.5.1	Die Notwendigkeit seriöser Quellen . . . . .	51
3.6	Die Herausforderung: Wahrheit und Wissen . . . . .	52
3.6.1	Der wissenschaftliche Prozess . . . . .	52
3.6.2	Die offene Frage: Der Umgang mit alternativen Lösungen . . . . .	52
3.7	Beispiele zur Veranschaulichung . . . . .	53
3.7.1	Die Macht des Chats . . . . .	53
3.7.2	Kollaborative Intelligenz . . . . .	54
3.8	Herausforderungen und Grenzen aktueller KI-Modelle . . . . .	54
3.8.1	Einstellbare Konversationsstile . . . . .	54
3.8.2	Fragen jenseits von Wikipedia . . . . .	54
3.8.3	Zukünftige Herausforderungen . . . . .	55
3.9	Aktuelle Grenzen und zukünftige Möglichkeiten . . . . .	55
<b>4</b>	<b>LLM für Sprache</b>	<b>57</b>
4.1	Begrüßung und aktueller Stand der AI-Technologie . . . . .	57
4.2	Generative AI und AI-Characters . . . . .	57
4.2.1	Übersetzungsleistung als Beispiel für semantisches Verständnis . . . . .	58
4.3	Herausforderungen und Erwartungen an zukünftige AI-Modelle . . . . .	58
4.3.1	Halluzination als Defizit aktueller Modelle . . . . .	58
4.4	Kompetenzbereiche aktueller und zukünftiger AI-Modelle . . . . .	59
4.4.1	Sprachkompetenz als Basis . . . . .	59
4.4.2	Erweiterbarkeit durch Kontextinformationen . . . . .	59
4.4.3	Bedeutung von Handlungsanweisungen . . . . .	59

4.5	Instruktionsausführung in der AI . . . . .	60
4.6	Lernen von Kompetenz in der AI . . . . .	60
4.6.1	Beispiel: Leonhard Euler . . . . .	61
4.6.2	Weitere Lernmöglichkeiten . . . . .	61
4.6.3	Digitalisierung historischer Bestände . . . . .	62
4.7	Generierung und Kontext in der Interaktion mit Chatmodellen . . . . .	62
4.8	Grenzen aktueller AI-Modelle . . . . .	63
4.9	Erwartungen an eine philosophische AI . . . . .	63
4.9.1	Allgemeine künstliche Intelligenz . . . . .	64
4.10	Semantische Suchen . . . . .	64
4.11	Reasoning . . . . .	65
4.11.1	Charakteristika eines AI-Modells mit Individualität . . . . .	66
4.12	Historische Vorbilder und Metaphern . . . . .	66
4.12.1	Der vitruvische Mensch - Proportion und Harmonie . . . . .	66
4.12.2	David - Freiheit und Selbstbestimmung . . . . .	68
4.12.3	Beuys und der tote Hase - Erklärung und Rechtfertigung . . . . .	70
4.13	Magister AI Faustus - Ein Arbeitstitel für die Zukunft . . . . .	71
4.13.1	Die Zusammenarbeit mit der Klassikstiftung Weimar . . . . .	73
4.13.2	Das Projekt: Goethes Biografie als Herausforderung für AI-Systeme . . . . .	73
4.13.3	Die Vision: Ein erweitertes AI-Modell . . . . .	74
4.14	Die Komplexität der Goethe-Quellen . . . . .	74
4.14.1	Die epistemische Herausforderung . . . . .	75
4.15	Organisation des Projekts . . . . .	75
4.16	Ausblick auf die kommenden Vorlesungen . . . . .	76
<b>5</b>	<b>Sprache und Text</b>	<b>77</b>
5.1	Rückblick auf die letzte Vorlesung . . . . .	77
5.1.1	Vorlesungsmanuskript durch AI generiert . . . . .	77
5.2	Das Projekt "Magister AI Faustus" . . . . .	78
5.2.1	Organisation des Projekts . . . . .	78
5.3	Entwurf einer philosophisch fundierten AI . . . . .	78
5.3.1	Kooperation mit der Klassik Stiftung Weimar . . . . .	79
5.4	Web-Interface unseres AI-Modells . . . . .	79
5.4.1	Logische Beziehungen zwischen Sätzen . . . . .	79
5.4.2	Defizite in logischen Schlussfolgerungen . . . . .	80

5.4.3	Sprachliche Anpassungen ohne Verbesserung . . . . .	81
5.4.4	Lichtblicke und ethische Bedenken . . . . .	81
5.4.5	Notwendigkeit eigener Definitionen . . . . .	81
5.5	Kompetenzen und Grenzen aktueller Modelle . . . . .	81
5.6	Zusammenfassungen generieren . . . . .	82
5.6.1	Technische Details . . . . .	82
5.7	Frage-Antwort-Dialoge . . . . .	83
5.8	Charakteristika und Fähigkeiten der Modelle . . . . .	83
5.8.1	Antwortformate . . . . .	83
5.8.2	Schreibstil . . . . .	83
5.8.3	Fachterminologie . . . . .	84
5.8.4	Kontextbezüge . . . . .	84
5.9	Aktuell bestehende Defizite . . . . .	84
5.9.1	Faktenwissen und logisches Denken . . . . .	84
5.9.2	Hermeneutik und Interpretation . . . . .	84
5.9.3	Kritik und ethische Bewertung . . . . .	85
5.10	Ausblick . . . . .	85
5.10.1	Bewertung der Teilaussagen und Korrekturbedarf . . . . .	86
5.11	Ein philosophisch fundiertes Handlungsmodell für den Umgang mit Instruktionen . . . . .	86
5.11.1	Die zwei Komponenten einer Instruktion als Handlungsanweisung . . . . .	87
5.11.2	Schritte des Handlungsmodells . . . . .	87
<b>6</b>	<b>Denken mit Logik</b>	<b>88</b>
6.1	Begrüßung und Einführung in die 6. Vorlesung . . . . .	88
6.1.1	Stärken und Schwächen der AI-Modelle . . . . .	88
6.1.2	Grenzen der AI-Modelle . . . . .	89
6.1.3	Einführung in das Projekt MAGISTER AI Faustus . . . . .	89
6.2	Logisches Denken . . . . .	89
6.2.1	Definition von logischem Denken . . . . .	90
6.2.2	Probleme beim Training von logischem Denken in AI-Modellen . . . . .	90
6.2.3	Unterschied zwischen Context of Discovery und Context of Justification . . . . .	91
6.2.4	Mangel an Discovery-Prozessen in Publikationen . . . . .	91
6.2.5	Experimentelle Untersuchung von Ideen während der Forschung . . . . .	93
6.3	Folgen für AI-Modelle . . . . .	94
6.3.1	Begrenzte Kompetenzen der AI-Modelle . . . . .	95

6.4 Instruktionen für AI-Modelle . . . . .	95
6.4.1 Formulierung von Forschungsvorhaben . . . . .	95
6.4.2 Das Lettre AI Studio . . . . .	96
6.5 Defizite der AI-Modelle bei logischen Verhältnissen . . . . .	96
6.5.1 Zensur bei Anthropic . . . . .	96
6.5.2 OpenAI hat dazugelernt . . . . .	97
6.5.3 Analyse der Sätze durch das einfache Modell . . . . .	97
6.5.4 Konfusion philosophischer Grundfähigkeiten . . . . .	98
6.5.5 Die Anfrage und das Scheitern des Modells . . . . .	99
6.5.6 Verbesserungen und Anpassungen der Modelle . . . . .	99
6.5.7 Die Bedeutung eigener Tests und Erfahrungen . . . . .	99
6.6 Analyse eines verbesserten Modells . . . . .	100
6.6.1 Korrekte Aussagen und logische Verhältnisse . . . . .	100
6.6.2 Problematische Feststellungen und Sachfragen . . . . .	100
6.6.3 Missverständnisse und fehlende Bezüge zur Frage . . . . .	100
6.7 Verbesserung durch Interaktion und Korrektur . . . . .	100
6.7.1 Lernfähigkeit und Grundlagenrevision . . . . .	101
6.8 Ausblick: Philosophie lehrt KI richtiges Denken . . . . .	101
6.9 Die Bedeutung der Schlüssigkeit . . . . .	102
6.9.1 Die Herausforderung der inhaltlichen Suche . . . . .	102
6.10 Die Grenzen aktueller KI-Modelle . . . . .	102
6.10.1 Linguistische Resolution als Lösungsansatz . . . . .	102
6.11 Ein praktisches Beispiel . . . . .	102
6.11.1 Analyse des Arguments . . . . .	103
6.11.2 Das Wahrheitswerttafelverfahren . . . . .	103
6.12 Die Macht der erweiterten Instruktionen . . . . .	103
6.13 Fazit . . . . .	104
<b>References</b>	<b>105</b>
<b>LettreAI Studio</b>	<b>106</b>

# **Philosophie der AI**

Diese Website enthält die Living Pages zur Vorlesung *Philosophie der KI* von Prof. Dr. Gerd Graßhoff. Die Vorlesung findet im Sommersemester 2024 an der Humboldt-Universität zu Berlin statt. Die Living Pages basieren auf dem Transkript der mündlich gehaltenen Vorlesung, das mit den Modellen von *Lettre AI* transkribiert und bearbeitet wurde.

Der transkribierte Inhalt wird von Gerd Graßhoff weiter redigiert, ergänzt und mit zusätzlichen Links und Verweisen angereichert.

Die Seiten richten sich an alle, die sich für die Philosophie der Künstlichen Intelligenz interessieren. Sie stehen unter der Creative Commons Lizenz 4.0 und dürfen gerne zitiert werden. Andere Nutzungen, wie auszugsweise Kopien oder digitale Verwendungen, erfordern die Erlaubnis des Autors.

# **1 Was ist AI?**

## **1.1 Begrüßung und Einführung**

Herzlich willkommen zur ersten Vorlesung “Philosophie der AI”! Ursprünglich trug diese Veranstaltung den Titel “Philosophie der künstlichen Intelligenz”, doch angesichts der aktuellen Diskussionen habe ich mich entschieden, den Begriff auf “AI” zu verkürzen. In diesem Semester möchte ich Ihnen einen umfassenden Überblick über die philosophischen Beiträge und Fundamente der modernen Artificial Intelligence geben und Sie durch die Grundlagen führen.

Entgegen der Erwartungen vieler geht es in dieser Vorlesung nicht primär darum, eine Bewertung oder Reflexion über die Folgen und Konsequenzen der künstlichen Intelligenz vorzunehmen. Obwohl wir diese Themen en passant ebenfalls behandeln werden, liegt der Kern der Vorlesung in der Erörterung der Grundthese, dass die eigentliche Innovation und der technologische Kern hinter dem Funktionieren der AI nicht nur in der Informatik, Technologie oder der fortschreitenden Entwicklung der Chips liegt, sondern in der Philosophie selbst. Ich vertrete die Ansicht, dass die künstliche Intelligenz heute eine Renaissance der analytischen Philosophie zur Folge hat, die die eigentliche inhaltliche und systematische Basis dessen bildet, was wir heute unter AI verstehen. Es handelt sich hierbei um eine anspruchsvolle Position, die die Philosophie nicht nur als Kommentator der technologischen und gesellschaftlichen Entwicklungen betrachtet, sondern als essenziellen Teil dieser Bewegung und Entwicklung.

Wir befinden uns derzeit nicht nur inmitten einer technologisch-gesellschaftlichen, politischen und sonstigen Revolution, die in ihrer Tragweite mit der Einführung der Elektrizität vor 150 Jahren oder des Webs vor etwa 25 Jahren vergleichbar ist. Vielmehr stehen wir gerade am Anfang einer Phase der technologischen Revolution durch die Einführung der künstlichen Intelligenz, deren weitreichende Entwicklungen wir nur erahnen können. Ein Indiz dafür ist die Tatsache, dass technologische Veränderungen, Möglichkeiten und Nutzungsformen mittlerweile auf täglicher Basis geschehen.

Während der Vorbereitung dieser Vorlesung ist mir aufgefallen, dass man nicht davon ausgehen kann, mit denselben Utensilien, Tools und Hilfsmitteln zu beginnen und am Ende der Vorlesung weiterzuarbeiten. Die Möglichkeiten und technologischen Anforderungen ändern sich so rasant, dass sie sich sogar

während des Verlaufs dieser Vorlesung weiterentwickeln werden. Mein Ziel ist es, Ihnen die Gelegenheit zu bieten, einige dieser Tools während der Vorlesung, in der Nachbereitung oder Vorbereitung selbst auszuprobieren.

Künstliche Intelligenz, oder kurz AI, ist ein Begriff für eine technische Möglichkeit, die Mitte der 50er Jahre die Phantasie einer Reihe von Forschern anregte.<sup>1</sup> Diese Phantasien entwuchsen den Arbeiten zu den Grundlagen der Mathematik und Logik, die eine enge Verwandschaft von zahlentheoretischen Fragestellungen mit denen von Algorithmen und der Berechenbarkeit von Problemen betrafen. Alan Turings Arbeiten als Fortsetzung von Kurt Gödels fundamentaler Arbeit über “unentscheidbare Sätze der Principia Mathematica und verwandter Systeme” war der Katalysator für die nachfolgenden Anstrengungen, die theoretischen Möglichkeiten in praktische Anwendungen zu überführen.<sup>2</sup> Ihr Ziel war es, maschinelle Computertechnologien zu entwickeln, die den menschlichen kognitiven Fähigkeiten nicht nur ebenbürtig sind, sondern sie sogar übertreffen. Man versprach damals vollmundig, dass dieses ehrgeizige Ziel in nur drei bis vier Jahren erreicht sein würde. Die Menschheit könnte dann endlich ihre Freizeit in vollen Zügen genießen, nur noch wenige Stunden pro Woche arbeiten, während der Rest von der AI erledigt würde.

Doch wie wir alle wissen, hat sich von dieser Vision bisher nichts eingelöst. Die Vorstellung war, dass AI als Meisterdisziplin des menschlichen Denkens schnell alle Bereiche überflügeln würde. Als Paradebeispiel galt damals das Schachspiel.<sup>3</sup> Doch erst Anfang der 2000er Jahre gelang es einem Computerprogramm, den Schachweltmeister Garri Kasparov in einem ernsthaften Spiel zu besiegen - immerhin 50 Jahre später als ursprünglich prophezeit.

Das andere große Ziel, Computer zu entwickeln, die selbstständig wissenschaftlich kreativ denken können, ist bis heute nicht wirklich erreicht. Trotz aller anderslautenden, manchmal sensationsheischenden Meldungen bin ich jedoch sicher, dass diese Stufe in den nächsten Jahren erreicht werden wird. Dass also wissenschaftliche, kreative, kognitive und intellektuelle Aktivitäten von Maschinen alleine, ohne Assistenz von Forschern gemeistert werden. Das ist sozusagen noch die Krönung der Herausforderung von AI, von Artificial Intelligence.

---

<sup>1</sup>[1], Dartmouth Summer Research Project, abgerufen am 15.5.2024.

<sup>2</sup>Turing skizzierte die Grundzüge eines universellen Computers in seiner Vorlesung in der London Mathematical Society 20. Feb 1947. [1], S. 378-394. [2], [3]. Von Neumann war tief beeinflusst von Turings Arbeit und setzte sie in der Entwicklung des EDVAC um. [1], S. 515.

<sup>3</sup>[4], [5], [6], [7], [8]

## 1.2 AI als Alleskönner

Was Ihnen derzeit tagtäglich in der Öffentlichkeit als AI präsentiert wird, hat mit den eigentlichen Visionen und Zielen oft wenig zu tun. Nehmen wir als Beispiel eine Anzeige der Firma Samsung für ihre "Bespoke AI 11-Kilogramm-Washing-Maschine Serie 8 mit AI-Eco-Bubble und Quick-Drive". Technisch gesehen handelt es sich schlicht um eine Waschmaschine, aber das Label "AI Wash" soll den Verkauf ankurbeln.



**Figure 1.1:** Samsung AI Waschmaschine

Was ist daran nun wirklich AI? Nicht viel, es ist mehr ein Verkaufsargument als alles andere. Alles, was halbwegs gesteuert ist, wird heutzutage als AI vermarktet. Wenn ich hier "Licht aus" sage und es dunkel würde, würden Sie vielleicht denken "Oh, wir haben AI an der HU". Dabei ist es letztlich nur eine etwas anspruchsvollere Steuerungstechnik, mehr nicht. Das Wort AI ist hier fehl am Platz, auch wenn es gerade überall en vogue ist.

### 1.2.1 Der Durchbruch der AI-Visionen

Sind wir also jetzt in einer Zeit angekommen, in der sich die ursprünglichen AI-Visionen doch noch erfüllen könnten? Meine Antwort lautet: Ja. Und ich möchte Ihnen heute einen systematischen Grund dafür nennen, der für mich entscheidend ist und den ich Ihnen so vermitteln möchte, dass er nachvollziehbar wird. Nebenbei bemerkt: Wenn Sie Fragen oder Zwischenfragen haben, melden Sie sich einfach. Dann gestalten wir die Vorlesung etwas lebendiger und interaktiver.

Der Aspekt, auf den ich hinaus möchte und den ich für den Meilenstein halte, ist, dass die AI-Visionen gerade dabei sind Wirklichkeit zu werden. Die AI-Propaganda hingegen, die sollten wir schnell beiseite legen. Das ist in erster Linie ein Verkaufsargument, das nicht den Kern der technologischen Innovation ausmacht. Und genau das soll heute unser Thema sein.

### **1.2.2 Die Attraktivität von AI**

Wo liegt denn potenziell die Attraktivität der AI, wie immer wir uns ihr auch nähern? Ist es eine bessere Internetsuchmaschine, die derzeit vielleicht eine der Triebfedern ist? Um das zu verstehen, müssen wir uns die Entwicklung des Internets vor Augen führen.

Gemessen an der Technologiegeschichte ist das Internet noch gar nicht so alt, etwas mehr als 20 Jahre. Wer die Anfänge noch miterlebt hat, erinnert sich an die ersten Browser, die damals oft mit Duschanlagen verwechselt wurden. Vor 20 Jahren wussten die wenigsten, was ein Internetbrowser eigentlich ist. Mittlerweile können wir uns ein Leben ohne Internet kaum noch vorstellen, weder technisch noch gesellschaftlich.

### **1.2.3 Die ursprüngliche Idee des Internets**

Im Kern war die Konstruktion des Internets, die am CERN entwickelt wurde, folgende: Irgendwo stellen wissenschaftliche Einrichtungen webzugängliche Seiten als Informationsquellen bereit. Als Wissenschaftler oder technologische Provider verantworten sie die Inhalte, pflegen sie und sorgen für dauerhafte Zugänglichkeit. Die Browser sind lediglich das lesende Frontend für diejenigen, die auf die Inhalte zugreifen wollen.

Damals war das Internet also eine Art anspruchsvolles Faxgerät als Empfänger der Inhalte. Der Clou lag darin, dass man ganz einfach andere Inhalte per Verlinkung einbinden konnte. So entwickelte sich ein Schneeballsystem, das ein globales Netz von miteinander verknüpften Inhalten erzeugte. Das war die Webrevolution vor 20 Jahren.

### **1.2.4 Die Ablösung der Webwelt durch AI**

Was wir jetzt erleben, ist eine Ablösung dieser Webwelt durch AI. In den nächsten Monaten werden Sie zunehmend feststellen, dass nicht mehr die Provider die Netzinhalte erstellen, auf Webservern bereitstellen und per Browser zugänglich machen. Diese Grundarchitektur wird abgelöst. Nicht mehr

der Browser verantwortet, pflegt und stellt die Inhalte bereit. Das ist eine revolutionäre Änderung der Architektur der Informationsflüsse, aber auch der damit verbundenen Probleme. Einen Teil davon werden wir noch kennenlernen oder haben Sie schon erfahren.

Das Web funktionierte bisher deshalb, weil die Inhalte von den jeweiligen Personen, Institutionen oder Wissenschaftlern, die sie bereitstellten, auch autorisiert wurden. Für die Korrektheit und Richtigkeit bürgten die Glaubwürdigkeit und Gewissenhaftigkeit der Provider. Das ändert sich jetzt. Und wir alle wissen um die Gefahren, aber auch Potenziale, die damit einhergehen.

- Auf der einen Seite sind es nun große Internetfirmen, die die Inhalte über AI-Maschinen, sogenannte Bots, bereitstellen.
- Auf der anderen Seite können es auch böswillige Gestalten, Institutionen oder Staaten sein, die Inhalte generieren, ins Netz einspeisen, ohne als autorisierende Internetprovider in Erscheinung zu treten.

Derzeit wird das unter dem Stichwort “Internetinhalte der Social Media” diskutiert. Doch das ist nur die Oberfläche. Der Kern des Wandels und des Problems liegt darin, dass die Grundarchitektur des Internets mit den verantwortlichen Providern abgelöst wird durch - ich will nicht sagen unverantwortliche Bots - aber zumindest durch nicht mehr verantwortliche Internetinhaltsprovider. Und das hängt eben mit der AI-Revolution und dem Wandel der Informationsflüsse im Internet zusammen.## Die Veränderung der Informationssuche im Zeitalter der Künstlichen Intelligenz

Meine sehr verehrten Damen und Herren, lassen Sie uns heute gemeinsam einen Blick in die Zukunft der Informationssuche werfen. Bislang war es für uns alle selbstverständlich, dass wir bei der Suche nach Informationen auf die Dienste von Suchmaschinen wie Google zurückgreifen konnten. Wir vertrauten darauf, dass die von diesen autoritativen Anbietern bereitgestellten Inhalte glaubwürdig und sorgsam kuriert waren. Doch in der nächsten Phase der digitalen Revolution wird sich dies grundlegend ändern.

### **1.2.5 Die Umgestaltung der Architektur des Internets**

Die Architektur des Internets befindet sich in einem extrem dynamischen Wandlungsprozess, dessen Ausgang noch niemand vorhersehen kann. Eines ist jedoch sicher: Es werden enorme Anstrengungen unternommen und gewaltige finanzielle Mittel investiert, um diese Transformation voranzutreiben. Jeder Staat, jede Region und auch Europa sollte ein vitales Interesse daran haben, die Kontrolle über diese Entwicklung nicht zu verlieren.

## 1.3 Neue Möglichkeiten durch Künstliche Intelligenz

Doch lassen Sie uns zunächst einen Blick auf die vielversprechenden Möglichkeiten werfen, die uns die Künstliche Intelligenz eröffnet. Vielleicht erscheinen Ihnen einige dieser Anwendungen auf den ersten Blick trivial, doch ich versichere Ihnen, sie haben das Potenzial, unseren Alltag und unsere Arbeit grundlegend zu verändern.

### 1.3.1 Hochwertige Übersetzungen

Nehmen wir zum Beispiel das Thema Übersetzungen. Seit Jahrzehnten wurden enorme Ressourcen in die Entwicklung von linguistischen Modellen zur automatischen Übersetzung von Sprachen investiert. Doch lange Zeit waren die Ergebnisse bestenfalls als Partygags zu gebrauchen und keinesfalls für den ernsthaften Einsatz geeignet. In den letzten Jahren hat sich dies jedoch grundlegend geändert. Mittlerweile sind die automatischen Übersetzungen von so hoher Qualität, dass sie sogar für akademische Zwecke genutzt werden können.

Lassen Sie mich Ihnen ein Beispiel aus meinem eigenen Fachgebiet, der Wissenschaftsgeschichte, geben. Viele der historischen Quellen, mit denen wir arbeiten, sind in Latein verfasst. Vor 100 Jahren mussten Doktoranden ihre Dissertationen an unserer Fakultät noch auf Latein einreichen. Heute würden die meisten von Ihnen wohl Schwierigkeiten haben, einen lateinischen Quelltext sinnvoll zu interpretieren. Doch dank der Fortschritte in der Künstlichen Intelligenz gibt es Hoffnung. Vielleicht führen wir ja in unserer Fakultät bald wieder die Pflicht ein, Doktorarbeiten auf Latein zu verfassen - mit AI als Hilfsmittel könnte dies durchaus ein Alleinstellungsmerkmal unserer Universität werden.

### 1.3.2 Simultanübersetzung und Lektoratsassistenten

Die Möglichkeiten gehen jedoch noch weiter. In naher Zukunft werden wir in der Lage sein, hervorragende Simultanübersetzungen anzubieten. Ausländische Studierende, die keine europäische Sprache beherrschen, könnten meine Vorlesung mit einem Ohrhörer verfolgen und eine simultane Übersetzung erhalten.

Auch im Bereich des Lektorats gibt es spannende Entwicklungen. Programme wie Grammarly oder DeepL Write bieten bereits heute Textverbesserungsvorschläge, die durchaus mit der Qualität professioneller Lektoratsassistenzen mithalten können. Selbst große wissenschaftliche Verlage wie Nature stellen ihren Autoren mittlerweile Tools zur Verfügung, um ihre englischen Texte in lesbare Form zu bringen. Ob und

wie dies gewünscht ist, wird derzeit heiß diskutiert. Doch ich bin davon überzeugt, dass in Zukunft das AI-gestützte Lektorat für wissenschaftliche Publikationen zum Standard werden wird.

### **1.3.3 Automatisierte Forschungsberichte**

Vor der Tür stehen bereits Modelle, die in der Lage sind, eigenständig Texte wie Forschungsberichte zu verfassen. In experimentellen Wissenschaften wie der klinischen Forschung wird bereits daran gearbeitet, Ergebnisse und Erkenntnisse automatisch in Berichte zu überführen, die qualitativ den gängigen Publikationen entsprechen. Dies wirft natürlich Fragen auf:

- Wer ist der Autor eines solchen Berichts?
- Akzeptieren wissenschaftliche Journals Texte, die von einer AI erstellt wurden?
- Wie gehen wir mit Verantwortlichkeit, Seriosität und Zurechenbarkeit um?

Diese Probleme müssen gelöst werden, wenn wir diese Entwicklung weiter vorantreiben wollen.

### **1.3.4 Das Labor *Lettre AI***

In eigenen Labor - *Lettre AI* (französisch für belesen, gebildet) erforschen und entwickeln wir die hier vorgestellten Techniken weiter. Unser Ziel ist es, eine AI weit über LLMs hinaus zu entwickeln, die auf der Basis der Fähigkeiten des Lesens, Übersetzens und Formulierens epistemische Qualifikationen mitbringt - also wissenbezogene Fähigkeiten, wie argumentieren, kritisieren, Evidenz nachweisen, Literatur vergleichen und aus einem Scholarium als Fundus wissenschaftlicher und kultureller Quellen schöpfen kann.

Lassen Sie mich Ihnen ein Beispiel für die bereits existierende Leistungsfähigkeit von AI geben. Ich zeige Ihnen hier einen Ausschnitt aus einem Werk, das zu Beginn des 17. Jahrhunderts wie ein Wirbelwind durch Europa fegte: den “Sidereus Nuncius” von niemand geringerem als Galileo Galilei. Dieses Buch markierte den Beginn einer Revolution, denn es war eines der ersten wissenschaftlichen Werke, das nicht nur auf Latein, sondern auch in der Volkssprache Italienisch verfasst wurde und so einer breiteren Öffentlichkeit zugänglich war.

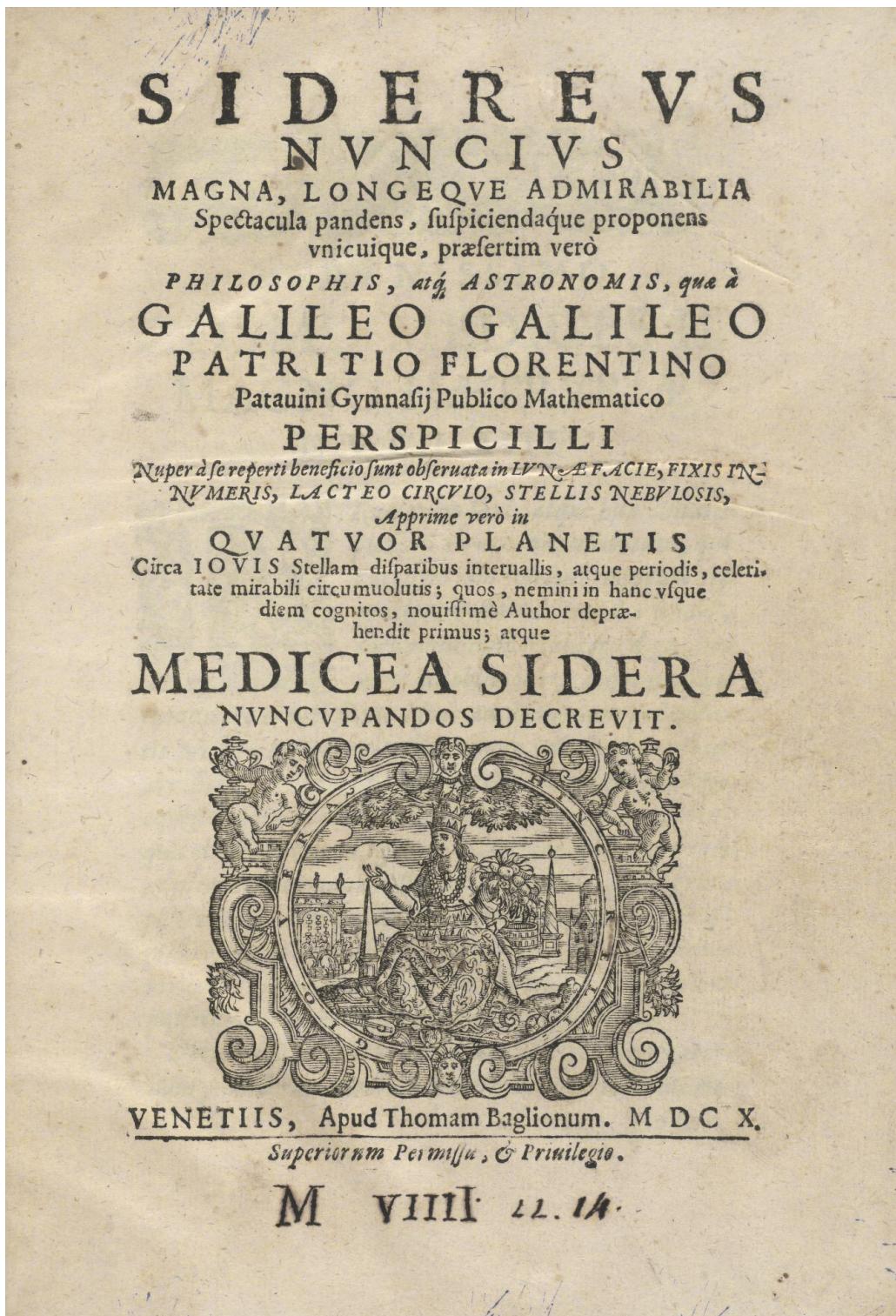


Figure 1.2: Sidereus Nuncius

Ich habe jetzt eine Variante von Chat-GPT aufgebaut. Für diejenigen unter Ihnen, die bereits mit Chat-GPT gearbeitet haben, wird die Oberfläche vertraut aussehen.

### **1.3.5 Übertragen eines Bildes in maschinenlesbaren Text**

Nehmen wir an, Sie haben eine Seite mit komplexen Inhalten vor sich, mit denen Sie in ihrer jetzigen Form nichts anfangen können. Hier kommt die AI ins Spiel: Sie können einfach einen Screenshot der Seite machen und diesen in den Chat-GPT hochladen. Anschließend instruieren Sie die AI mit einer Anweisung wie “Transkribiere das Bild” - und schon erhalten Sie eine nahezu fehlerfreie Übertragung des nicht gerade einfachen Textes in getippte Buchstaben. Eine Leistung, die bis heute kein anderes Programm in dieser Qualität vollbringen kann.

### **1.3.6 Übersetzen des Textes in eine andere Sprache**

Doch das ist erst der Anfang. Nehmen wir an, Sie verstehen kein Latein - kein Problem. Tippen Sie einfach “Übersetze diesen Text ins Deutsche” ein und schon erhalten Sie eine verständliche, wenn auch noch etwas gewöhnungsbedürftige Übersetzung. Mit ein wenig Feinschliff oder dem Wechsel des Modells lässt sich daraus ein publikationsreifer deutscher Text erstellen. Und das Ganze funktioniert nicht nur für Deutsch und Englisch, sondern für über 150 Sprachen weltweit, darunter auch Japanisch und Koreanisch. Selbst obskure mittelalterliche Quellen stellen kein Hindernis dar.

## **1.4 Erweiterungen**

Doch jetzt fängt der eigentliche Spaß erst an. Mit dem nun zugänglichen Text eröffnen sich ganz neue Möglichkeiten jenseits der typischen Google-Fragen wie “Wer war Galilei?” oder “Wann lebte er?”. Stattdessen können Sie die AI mit Fragen herausfordern, die Google unmöglich beantworten kann. Zum Beispiel: “In welcher Stadt trank Galilei im Mai 1615 ein Glas Wein?”. Das Problem liegt hier nicht nur darin, dass Google dieses spezifische Ereignis nicht kennt, sondern dass eine einfache Stichwortsuche prinzipiell nicht ausreicht, um die Antwort zu finden.

### **1.4.1 Analogie zu Sherlock Holmes**

Stellen Sie sich die AI als eine Art elektronischen Sherlock Holmes vor. Sie nimmt das gesamte Universum an Dokumenten über Galilei zur Kenntnis - seine Briefe, seine historischen Lebensumstände, seine typischen Aktivitäten im Frühjahr 1605. Aus diesen Informationen zieht sie dann Rückschlüsse und generiert eine fundierte Hypothese darüber, wo und wann Galilei wahrscheinlich sein Glas Wein genossen hat. Zwar nicht mit absoluter Sicherheit, aber basierend auf seinen regelmäßigen Lebensumständen. Solche Fragen werden die AI-Modelle in naher Zukunft beantworten können.

### **1.4.2 Vielfältige Analysemöglichkeiten von Texten**

Doch damit nicht genug. Sie können die AI auch anweisen, eine Tabelle mit allen Verben des Textes zu erstellen oder gezielt nach Verben zu suchen, die ein Lob, eine Ankündigung oder ein Versprechen ausdrücken - selbst wenn Sie die genaue Formulierung nicht kennen. Die Möglichkeiten sind schier grenzenlos.

Ein konkretes Beispiel: Fragen wir die AI, wer sich laut dem Text bewegt. Nach kurzer Bedenkzeit liefert sie die korrekte Antwort: Die vier Planeten bewegen sich zu verschiedenen Zeiten und mit erstaunlicher Geschwindigkeit um den Stern Jupiter - eine Entdeckung, die Galilei machte und die tatsächlich im lateinischen Originaltext erwähnt wird.

## **1.5 Philosophie als Grundlage für die Möglichkeiten der AI**

Doch wie ist das alles möglich? Die Antwort liegt in der Philosophie - nicht in der Technik. Natürlich brauchen wir auch die technische Infrastruktur, so wie wir Beamer und Notebooks benötigen. Aber der eigentliche Schlüssel zu den Fähigkeiten der AI ist philosophischer Natur. Das wird oft übersehen, doch ich möchte Ihnen zeigen, warum Philosophie hier so entscheidend ist.

### **1.5.1 Beantwortung von Fragen über Mikrofoneingabe**

Um das Potenzial der AI weiter zu verdeutlichen, können wir auch das Mikrofon aktivieren und eine Frage stellen: "Hat Galilei diese Entdeckung selbst durch Beobachtungen gemacht?". Das System denkt kurz nach und liefert dann die zutreffende Antwort: Ja, laut den Angaben im Text hat Galilei die Entdeckung tatsächlich selbst durch Beobachtungen gemacht.

Das Erstaunliche daran ist nicht nur, dass überhaupt eine Antwort generiert wird, sondern vor allem die Qualität dieser Antwort - trotz Versprechern und spontaner Formulierung meinerseits.## Einführung in die sprachliche Dimension der AI

Meine Damen und Herren, heute möchte ich Ihnen eine faszinierende und zugleich beunruhigende Entwicklung in der Welt der künstlichen Intelligenz näherbringen. Es geht um die Fähigkeit von AI-Systemen, nicht nur Informationen aus autoritativen Quellen zu sammeln, sondern eigenständig Antworten zu generieren und Inhalte zu erstellen. Diese Entwicklung hat weitreichende Konsequenzen für unser Verständnis von Wissen und Informationsverarbeitung.

### **1.5.2 Die Möglichkeiten der AI**

Die Möglichkeiten der AI sind atemberaubend und erweitern sich täglich. Lassen Sie mich Ihnen einige Beispiele nennen:

- Übersetzung: AI-Systeme können Texte von einer Sprache in eine andere übersetzen, und zwar mit einer Genauigkeit und Geschwindigkeit, die menschliche Übersetzer in den Schatten stellt.
- Bild-zu-Text-Konvertierung: AI kann Bilder analysieren und deren Inhalt in Textform beschreiben. Dies eröffnet völlig neue Möglichkeiten der Bildverarbeitung und -archivierung.
- Audio-zu-Text-Konvertierung: Gesprochene Sprache kann von AI-Systemen in Echtzeit transkribiert werden, was die Erstellung von Protokollen und Untertiteln erleichtert.
- Textzusammenfassung: Geben Sie der AI ein ganzes Buch, und sie wird Ihnen eine prägnante Zusammenfassung liefern. Dies kann die Recherche und das Studium enorm beschleunigen.
- Text-zu-Audio-Konvertierung: Umgekehrt kann AI auch geschriebenen Text in gesprochene Sprache umwandeln, was neue Möglichkeiten für Hörbücher und Sprachassistenten eröffnet.
- Text-zu-Video-Konvertierung: Hier wird es geradezu unheimlich. AI kann aus Textbeschreibungen realistische Videos generieren, die kaum noch von echten Aufnahmen zu unterscheiden sind.

### **1.5.3 Gefahren der AI**

So faszinierend diese Möglichkeiten auch sind, sie bergen auch erhebliche Risiken. Ein zentrales Problem ist das Phänomen der “Halluzination”. Dabei generiert die AI scheinbar plausible Informationen, die jedoch nicht der Realität entsprechen.

Ein Beispiel: Ich fragte eine AI nach dem Namen der zweiten Frau des Mathematikers Leonhard Euler. Die Antwort klang überzeugend, inklusive eines Verweises auf eine Publikation der Petersburger Akademieschriften von 1784. Doch diese Publikation existiert gar nicht, und die genannte Person war nie mit Euler verheiratet.

Solche Halluzinationen können fatale Folgen haben, wenn sie unerkannt bleiben. Wer eine solche Information zitiert, disqualifiziert sich wissenschaftlich für immer. Dieses Problem trat auch bei der Mars-Mission der NASA auf, als eine AI falsche Informationen über einen Erkundungssatelliten verbreitete.

#### **1.5.4 Der sprachliche Kern der AI**

Bei all diesen Anwendungen, sei es Bild-, Audio- oder Videoverarbeitung, bildet die Sprache den Kern der AI-Technologie. Selbst bei der Bildanalyse übersetzt die AI zunächst das Bild in eine verbale Beschreibung, bevor sie weiterverarbeitet wird.

Diese Erkenntnis ist philosophisch bedeutsam und erinnert an Wittgensteins These von der Unhintergehbarmkeit der Sprache. Die sprachliche Verbalisierung von Inhalten ist der Dreh- und Angelpunkt der AI, und genau darum soll es in dieser Vorlesung gehen.

Ich werde mich nicht auf die technischen Details der AI-Entwicklung konzentrieren, sondern auf den Umgang mit Sprache in AI-Modellen. Die anderen Medien sind zwar faszinierend, aber letztlich sekundär. Unser roter Faden wird die philosophische Dimension der sprachlichen Verarbeitung in der AI sein.## Gefahren und Probleme der künstlichen Intelligenz

Meine Damen und Herren, lassen Sie uns heute über die Schattenseiten der künstlichen Intelligenz sprechen. Wir haben bereits die atemberaubenden Möglichkeiten dieser Technologie gesehen, doch nun ist es an der Zeit, auch die Probleme und Gefahren zu beleuchten, die damit einhergehen.

#### **1.5.5 Das Problem der Halluzinationen**

Eines der ersten Probleme, auf das wir stoßen, sind die sogenannten Halluzinationen der AI-Modelle. Ein eindrucksvolles Beispiel dafür lieferte das Supermodell von Google, das auf die Frage "Wer fliegt denn da?" eine Antwort gab, die zwar plausibel klang, aber rein fiktiv war. Ohne Zugriff auf aktuelle NASA-Informationen oder Tagesnachrichten erfand das Modell kurzerhand einen Satellitennamen. Innerhalb einer halben Stunde wurde es vom Netz genommen, und der Marktwert von Google-Aktien sank um Millionen. Seitdem trauen sich die Unternehmen nicht mehr, ihre Modelle zu veröffentlichen.

Doch warum halluzinieren die Modelle überhaupt, wenn sie doch schon so viele Fähigkeiten besitzen? Die Antwort darauf ist komplexer als man denkt.

### **1.5.6 Die Gefahr der Manipulation durch glaubwürdige Fakes**

Ein weiteres Problem, das eng mit den Halluzinationen verbunden ist, ist die Fähigkeit der AI, glaubwürdige Texte, Bilder und sogar Videos zu produzieren. Dies öffnet Tür und Tor für falsche oder manipulative Informationen, die auf den ersten Blick echt erscheinen.

Ein aktuelles Beispiel dafür sind die Videos, die im Zusammenhang mit dem Raketenüberfall auf Israel in den sozialen Medien aufgetaucht sind. Sie zeigten panische Einwohner von Tel Aviv, die vor nicht existierenden Einschlägen flohen. Diese Videos wurden absichtlich generiert, um die Öffentlichkeit zu täuschen, und sind für den Betrachter zunächst nicht als Manipulation zu erkennen.

### **1.5.7 Selektive Informationen und die Pluralität der Hintergründe**

Jede Antwort, die uns ein AI-Modell gibt, basiert auf bestimmten Annahmen und Voraussetzungen. Diese haben jedoch immer auch Alternativen, die möglicherweise nicht besser oder schlechter sind, aber eine Pluralität an Hintergründen darstellen.

Wenn wir eine bestimmte Antwort akzeptieren, akzeptieren wir auch die Voraussetzungen dafür und vernachlässigen die Alternativen. Ein Beispiel dafür ist die Anfrage an ein AI-Modell, ein Porträt eines möglichen Nachfolgers des jetzigen Papstes zu erstellen. Aufgrund der politisch korrekten Voreinstellung des Modells wurde eine farbige Frau im Papstgewand generiert - eine Darstellung, die in der Realität aufgrund der Zusammensetzung des Kardinalskollegiums höchst unwahrscheinlich ist.

Dieses Beispiel verdeutlicht, wie selektive Informationen zu verzerrten Ergebnissen führen können. Es wirft die Frage auf, wie wir mit diesen Problemen umgehen sollen.

### **1.5.8 Die Unausweichlichkeit der AI-Entwicklung und die Notwendigkeit der Gestaltung**

Eines ist klar: Wir können uns vor diesen Fragen nicht drücken. Die Entwicklung der künstlichen Intelligenz ist unwiderstehlich und unausweichlich. Ab heute werden uns diese Technologien mit all ihren Vor- und Nachteilen zunehmend beschäftigen.

Wir müssen lernen, damit umzugehen und die Entwicklung aktiv mitzugestalten. Nicht im Sinne einer Kontrolle, sondern einer Gestaltung. Denn wenn wir jetzt nicht eingreifen, laufen wir Gefahr, die Kontrolle über diesen Prozess zu verlieren.

### **1.5.9 Weitere Gefahren: Diskriminierung und Überwachung**

Neben der selektiven Information gibt es weitere Gefahren, die wir im Auge behalten müssen. Dazu gehören Dimensionen der Diskriminierung, bei denen bestimmte Personengruppen oder Qualifikationen berücksichtigt werden, andere hingegen nicht.

Auch die Möglichkeiten der Überwachung durch AI-Systeme sind alarmierend. Ein Beispiel dafür ist China, wo Besucher bei der Einreise lediglich in eine Kamera lächeln müssen und dann während ihres gesamten Aufenthalts live verfolgt und protokolliert werden.

Diese Entwicklungen werfen Fragen auf, wie weit solche Technologien zugelassen und kontrolliert werden sollten. Eine Antwort darauf zu finden, ist keine leichte Aufgabe.

### **1.5.10 Die Notwendigkeit der Auseinandersetzung mit AI**

Angesichts dieser erschütternden Probleme könnte man geneigt sein, das Thema AI einfach zu vergessen. Wozu sich mit Übersetzungen von Galileis lateinischen Texten beschäftigen, wenn wir dafür doch unsere Gelehrten haben?

Doch so einfach ist es nicht. Die Vorteile der künstlichen Intelligenz sind zu groß, um sie zu ignorieren. Wir müssen uns mit dieser Technologie auseinandersetzen, ihre Möglichkeiten nutzen und gleichzeitig ihre Schattenseiten im Blick behalten. Nur so können wir eine Zukunft gestalten, in der die AI zum Wohle der Menschheit eingesetzt wird.

## **1.6 Nutzungsmöglichkeiten in der Wissenschaft**

Lassen Sie uns ein Beispiel betrachten, das wir gerade schon diskutiert haben. In der Fachliteratur hält sich hartnäckig das Gerücht, dass Galileis Vater sich negativ über die wissenschaftliche Nutzung anderer Sprachen als Latein geäußert haben soll. Das würde natürlich einen spannenden Vater-Sohn-Konflikt darstellen, denn Galilei selbst ist ja berühmt dafür, dass er das Italienische für die Wissenschaft nutzbar machte, indem er auf Italienisch publizierte.

In zahlreichen Sekundärquellen findet man die These, dass sein Vater dies nicht für wissenschaftlich hielt und dass sein Sohn Galileo Galilei sich besser von diesen italienischen Publikationen fernhalten sollte. Oh, Moment mal - da steht, dass Kepler sich gegenüber Galilei negativ geäußert hat, nicht Galileis Vater. Danke für den Hinweis! Das ist keine Halluzination, sondern ein echter Fehler meinerseits. Ich hoffe, ich vergesse nicht, das für die Internetversion zu korrigieren.

Die Pointe ist jedenfalls, dass man eine solche Frage - ob sich eine Person X irgendwo negativ zu einer bestimmten These geäußert hat - mit Google nicht beantworten kann. Das mag trivial klingen, aber im Moment ist es tatsächlich nicht möglich, dies durch eine Google-Suche herauszufinden. Warum? Weil Google Ihnen kein Dokument im Internet liefern wird, in dem diese Frage direkt beantwortet wird. Und wenn es ein solches Dokument nicht gibt, ist die Frage für Sie mit Google-Techniken nicht zu beantworten.

Dabei handelt es sich um eine Frage, die historisch gesehen entweder wahr oder falsch ist. Wie kann man das also entscheiden? Nicht mit den heutigen Google-Techniken. Hier braucht es eine neue Dimension der Recherche, die über bestimmte Fähigkeiten verfügen muss.

## 1.7 Bislang nicht lösbarer Aufgaben

Lassen Sie mich Ihnen anhand einer Liste von Aufgaben und Fragen veranschaulichen, wie zunehmend Probleme auftauchen, die mit den heutigen akademischen Techniken nicht zu lösen sind. Ich spreche hier von Fragen, die selbst Sie als forschende Person nicht beantworten können, wenn sie halbwegs komplex sind.

Mir geht es um die unlösbaren Probleme der realen Forschungswelt, die zwar mit AI lösbar wären, aber aufgrund bestimmter fehlender Fertigkeiten bisher nicht gelöst werden können. Jetzt befinden wir uns im philosophischen Teil meiner Ausführungen und ich werde versuchen, dies sprachanalytisch zu komprimieren.

### 1.7.1 Frage 1: Einfache Aussage in einer Quelle

Angenommen, Person A äußert sich in einer Quelle Q zu einer Person namens Jochen Schmidt. Ist diese Aussage wahr oder falsch? Hier haben Sie noch eine gewisse Chance, die Frage eindeutig zu beantworten, wenn Sie die Quelle Q gefunden haben und darin die Person A benannt wird und sich zu Jochen Schmidt äußert. Der Anforderungsgrad ist hier noch nicht sehr hoch. Wenn das Ihre Examensaufgabe wäre, hätten

Sie eine realistische Chance, sie zu lösen. Sie müssten nur so lange alle Quellen durchlesen, bis Sie die richtige gefunden haben.

### **1.7.2 Frage 2: Aussage in Briefen zu einem Thema**

Nehmen wir an, Person A äußert sich in ihren Briefen zu einem Thema T. Das können Sie schon nicht mehr ohne weiteres lösen, ohne eine Lebensdauer damit zu verbringen, das gesamte Schrifttum von Person A zu lesen. Wenn Sie z.B. für eine Examensarbeit eine Biografie über eine Person namens Heinz Müller verfassen sollten und eine solche Aufgabe hätten, müssten Sie zunächst alle Briefe zusammentragen und sie komplett lesen. Und selbst dann wären Sie sich nicht sicher, ob Sie wirklich alle Briefe gefunden haben.

Denken Sie nur an die Kafka-Forscher. Wenn Sie wissen wollen, ob sich Kafka in seinen Briefen jemals zu einem bestimmten Thema geäußert hat oder nicht, haben Sie einen enormen manuellen Forschungsaufwand vor sich, um überhaupt in die Nähe einer Antwort zu kommen. Hier befinden wir uns bereits in Bereichen, die schwer zu beantworten sind - Fragestellungen, die bislang praktisch nicht zu lösen waren.

### **1.7.3 Frage 3: Aussagen einer Person in ihren Schriften**

Hat eine Person A in ihren Schriften Aussagen der Art T getroffen, wenn Person A sehr viel geschrieben hat? Nehmen wir als Beispiel die Briefe Napoleons. Hat sich Napoleon jemals zu Aspekten der Vorläufer der Genfer Konvention bei der Kriegsführung geäußert? Das können Sie aus praktischen Gründen nicht lösen. Ich will an dieser Stelle nicht sagen, dass es prinzipiell unmöglich ist, aber in der Wissenschaft möchte man solche Fragen beantwortet haben. Und das gilt nicht nur für das öffentliche Interesse, sondern auch für die Wissenschaft selbst.

Sie können sich vorstellen, welch enorme Konsequenzen es für die Wissenschaft hätte, wenn man solche Fragen überhaupt beantworten könnte. Dann wäre es möglich, weitreichende Thesen zu Napoleons Verständnis von Krieg und Frieden aufzustellen, die von der Evidenz abhängen, mit der man solche Fragen beantworten kann. Im Moment ist das nicht möglich.

### **1.7.4 Frage 4: Keine Aussage einer Person in ihren Schriften**

Angenommen, Person A hat in ihren Schriften keine Aussage T getroffen. Als normaler arbeitender Historiker oder Geisteswissenschaftler werden Sie diese Frage nicht seriös beantworten können. Deshalb

gibt es in der Literatur die Unsitte, andere Werke zu zitieren, die sich aus irgendwelchen Gründen dazu bemüßigt fühlten, solche Fragen zu beantworten.

Ein Beispiel: Nehmen wir wieder Kafka. Manche Autoren vertreten die These, dass Kafka sich nie antisemitisch geäußert hat. Aber welche Evidenz können Sie dafür eigentlich angeben? Es ist schwierig, eine nicht vorhandene Lektüre von Briefen als Beleg anzuführen. Wie wollen Sie eine solche These rechtfertigen, wenn Sie sie vertreten?

Eine der größten Unsitten der gegenwärtigen akademischen Literatur besteht darin, nicht selbst das Risiko einer These einzugehen, sondern stattdessen den berühmten Heinz Müller zu zitieren, weil er schon einmal etwas Ähnliches gesagt hat. Also fügt man eine Fußnote in die Arbeit ein: "Heinz Müller, 1973, Seite 5: Ganz klar, Kafka hat sich nie antisemitisch geäußert." Und auf einmal entsteht ein Schneeballsystem, das dem Halluzinationseffekt ähnelt, den wir gerade hier hatten. Und zwar nur deshalb, weil die Evidenz, die für bestimmte Thesen erforderlich ist, auf manuelle Weise kaum zu beschaffen ist. Mit AI werden Sie das in Zukunft können.

## 1.8 Die Herausforderung der inhaltlichen Analyse mit AI

Jetzt werden Sie vielleicht fragen: Inwiefern ist das speziell für AI relevant? Man könnte doch erwarten, dass sich das grammatisch lösen lässt. Wenn ich die Aussage T formalisieren kann, müsste ich doch auf dem Textkorpus einfach prüfen können, ob diese Bedingung irgendwo erfüllt ist, oder?

Genau das ist der springende Punkt, und ich muss jetzt ein bisschen auf die Uhr schauen, damit ich meine Kurve hier noch hinbekomme. Aber diese Kurve berührt schon das Thema. Was heißt es, in Ihrem Korpus prüfen zu können?

Nehmen wir an, Sie hätten den Idealfall: Kafkas gesammelten Briefwechsel in einer Datenbank. Jetzt möchten Sie wissen, ob es darin eine antisemitische Formulierung gibt. Wie sieht die denn aus? Wenn Sie Ihre Datenbank nach Art einer Google-Suche nach bestimmten Wortvorkommnissen durchforsten, dann können Sie das lösen. Das ist die klassische Vorgehensweise.

Aber inhaltlich betrachtet: Was ist eigentlich eine antisemitische Äußerung? Sobald es darum geht - und deshalb habe ich es hier erwähnt - kön## Betrachtungen zur künstlichen Intelligenz und Sprachverarbeitung

Meine sehr geehrten Damen und Herren, liebe Studierende,

in der heutigen Vorlesung möchte ich Ihnen einen faszinierenden Einblick in die Welt der künstlichen Intelligenz und insbesondere deren Fähigkeiten zur Sprachverarbeitung geben. Wir werden uns mit der Frage

beschäftigen, inwieweit AI-Systeme in der Lage sind, komplexe sprachliche Konstrukte wie Metaphern, Ironie oder versteckte Bedeutungen zu erkennen und zu interpretieren.

### **1.8.1 Grenzen der traditionellen Datenbanken**

Zunächst einmal möchte ich klarstellen, dass ich keineswegs behauptet habe, es gäbe in den vorliegenden Dokumenten keine relevanten Satzvorkommnisse. Die herkömmliche Art der Dokumentenaufzeichnung und -abfrage, wie sie etwa mit Datenbanken möglich ist, erlaubt zwar das Auffinden bestimmter Textpassagen, jedoch keine inhaltlichen Suchen im eigentlichen Sinne.

Selbst moderne AI-Systeme können nicht mit absoluter Sicherheit feststellen, dass eine bestimmte Aussage nicht getroffen wurde, da stets die Möglichkeit besteht, dass die zugrunde liegende Datenbasis unvollständig ist. Vielmehr lässt sich hier nur mit Wahrscheinlichkeiten operieren - ein Begriff, den ich an dieser Stelle allerdings kritisch hinterfragen möchte.

### **1.8.2 Qualifizierte Aussagen auf Basis der verfügbaren Evidenz**

Wahrscheinlichkeiten sind numerische Werte zwischen 0 und 1, die man in diesem Kontext nicht sinnvoll einsetzen kann. Stattdessen sollte man sich auf die konkrete Situation beziehen und feststellen: Auf Basis dieser und jener Grundgesamtheit von Briefwechseln und Äußerungen, die als Dokumente für die Befunde zur Verfügung stehen, lässt sich unter der Voraussetzung, dass sie die alleinige Entscheidungsgrundlage bilden, folgendes Fazit ableiten.

Eine solche differenzierte Betrachtung der Befundlage ist unerlässlich, denn es lässt sich ja nicht ausschließen, dass genau jene Briefe, die möglicherweise relevante Inhalte enthalten, vernichtet wurden. Ein solches Szenario würde den Wahrheitswert der Fragestellung grundlegend verändern. Auch AI-Systeme können diese Problematik nicht vollständig ausräumen, sehr wohl aber eine qualifizierte, auf der verfügbaren Evidenz basierende Antwort geben.

### **1.8.3 Herausforderungen bei der Interpretation von Metaphern und Ironie**

Ein besonders spannendes Feld ist die Fähigkeit von AI-Systemen, mit Metaphern und uneigentlichem Sprachgebrauch umzugehen. Gerade im Kontext des Antisemitismus verbergen sich oft codierte Botschaften hinter scheinbar harmlosen Formulierungen. Während eine Blut-und-Boden-Ideologie

relativ leicht zu identifizieren ist, stellt die Interpretation von Begriffen wie “entwurzelt” oder “ohne Verwurzelung” eine ungleich größere Herausforderung dar.

Anhand eines konkreten Beispiels möchte ich Ihnen verdeutlichen, wozu moderne AI-Systeme in diesem Bereich bereits in der Lage sind. In München hatten wir es mit revolutionären Briefen aus der Zeit der Französischen Revolution zu tun, die in elegantem Französisch verfasst waren und vor Ironie und Sarkasmus nur so strotzten. Um diese Feinheiten zu erkennen, bedarf es zunächst einmal exzellenter Sprachkenntnisse. Doch selbst dann gilt es, die ironischen Komponenten als solche zu identifizieren.

Ich kann Ihnen versichern, dass AI-Systeme mittlerweile über eine Sprachkompetenz verfügen, die es ihnen erlaubt, auch diese Dimension der Sprachverwendung zu erkennen. Allerdings dürfen Sie sich das nicht als simples Schwarz-Weiß-Schema vorstellen, bei dem man einfach einen “Ironie-Kompetenz-Knopf” umlegt und schon funktioniert alles wie bei einem literarischen Meisterinterpret.

#### **1.8.4 Lernfähigkeit und Entwicklungspotenzial von AI-Systemen**

Vielmehr müssen Sie sich den Lernprozess der AI ähnlich vorstellen wie Ihre eigene Entwicklung zu Beginn Ihres Studiums. Auch Sie haben im Laufe der Zeit eine Menge dazugelernt und sich weiterentwickelt. Genauso können auch AI-Modelle lernen und sich verbessern. Ich möchte keineswegs behaupten, dass bereits alle Probleme und Herausforderungen gelöst sind, aber es gibt vielversprechende Lösungsansätze, um auch mit komplexeren Formen der Sprachverwendung umgehen zu können.

In München haben wir beispielsweise erfolgreich getestet, ob AI-Systeme in der Lage sind, bissige Karikaturen aus den 1920er Jahren zu interpretieren und zu erkennen, welche Personen mit welchen Klischees auf den Arm genommen werden. Mit dem richtigen Training ist es den Bilderkennungsalgorithmen tatsächlich gelungen, diese Zusammenhänge zu entschlüsseln.

#### **1.8.5 Der Paradigmenwechsel durch Large Language Models und Embeddings**

Der entscheidende Unterschied und gleichzeitig der Punkt, an dem der “Philosophical Turn” der AI einsetzt, liegt in der Entwicklung von Techniken wie Large Language Models oder Embeddings. Diese ermöglichen eine Abkehr von der reinen Textsuche hin zu einer inhaltlichen Erfassung der Bedeutung sprachlicher Ausdrücke. Dieser semantische Wechsel, den ich auch gerne als “Semantic Turn” bezeichne, ist der Schlüssel zu den beeindruckenden Fähigkeiten moderner AI-Systeme.

Egal ob es um die Analyse von Bildern, Texten oder Audioaufnahmen geht - all diesen Anwendungen liegt zugrunde, dass die Systeme nicht nur nach bestimmten Zeichenfolgen suchen, sondern deren Bedeutung

erfassen und identifizieren können. Genau darum geht es bei den milliardenschweren Investitionen in diesem Bereich: den Modellen beizubringen, auf Basis der eingegebenen Daten die dahinterstehende Semantik zu erkennen.

### **1.8.6 Die Bedeutung der Philosophie für die AI-Forschung**

Damit eröffnet sich ein weites Feld für die Philosophie. Solange wir nur von Sätzen sprechen, bewegen wir uns auf der Ebene von Formulierungen und syntaktischen Strukturen. Wenn wir jedoch nach der Bedeutung eines Ausdrucks fragen, betreten wir Neuland. Genau hier setzt die aktuelle AI-Revolution an, und deshalb ist die Philosophie von zentraler Bedeutung für diese Entwicklung.

Als Studierende der Philosophie sollten Sie mit der klassischen Unterscheidung zwischen Satz und Aussage vertraut sein. Im Deutschen ist diese Differenzierung von größter Wichtigkeit, während sie in englischen Übersetzungen oft vernachlässigt wird. So haben etwa die Übersetzer von Wittgensteins Gesammelten Werken sowohl für "Aussage" als auch für "Satz" durchgängig den Begriff "Sentence" verwendet, was zu erheblichen Missverständnissen führen kann. Im Englischen heißt es korrekterweise "Sentence" für Satz und "Proposition" für Aussage.

Genau diese Unterscheidung markiert die fundamentale Revolution, die sich gerade vollzieht: Wir haben es nun mit Maschinen zu tun, die mit Aussagen umgehen können. Und nur Aussagen, nicht Sätze, können wahr oder falsch sein. Wer also über Fake News, Halluzinationen und ähnliche Phänomene spricht und sich dabei auf Sätze bezieht, liegt philosophisch gesehen völlig falsch. Wahrheit und Falschheit können sich konzeptionell nur auf Aussagen beziehen.

Die Tatsache, dass AI-Systeme nun in der Lage sind, sich mit Aussagen zu befassen, birgt ebenso faszinierende Möglichkeiten wie Gefahren. In der nächsten Vorlesung werden wir uns eingehender mit diesen Aspekten beschäftigen und uns ansehen, wie genau diese neuen Technologien funktionieren und welche Auswirkungen sie haben können.

# **2 Die Revolution der AI**

## **2.1 Begrüßung und Rückblick auf die letzte Vorlesung**

Herzlich willkommen zur zweiten Vorlesung "Philosophie der AI"! Lassen Sie uns zunächst an die bemerkenswerten Leistungen der AI erinnern, von denen wir uns versprechen, dass sie auch in der geisteswissenschaftlichen Forschung etwas Außergewöhnliches hervorbringen können. Wir hoffen, dass die AI unser tägliches Forschungsgeschehen in den Geisteswissenschaften bereichern und erleichtern wird.

## **2.2 Traditionell schwer lösbarer Fragen in der Forschung**

In der Forschung und im Studium stoßen wir immer wieder auf Fragen, die zwar selbstverständlich erscheinen, aber dennoch eine Herausforderung darstellen. Ein Beispiel dafür ist die Suche nach Evidenz in Quellen innerhalb eines definierten Kreises von Texten und Fachbüchern, die ich als "Scholarium" bezeichne. Je nach Komplexität der historischen Aussage H kann der Nachweis solcher Evidenz sehr arbeitsintensiv sein. Dank der AI werden wir in Zukunft, abhängig von der Zugänglichkeit und Aufbereitung des Scholariums, solche Fragen schnell und mühelos beantworten können.

### **2.2.1 Noch schwieriger: Evidenz zur Widerlegung von Hypothesen finden**

Eine noch größere Herausforderung stellt die Suche nach Evidenz zur Widerlegung einer Hypothese H dar. Im wissenschaftlichen Alltag ist dies praktisch unmöglich, obwohl wir solche Aussagen häufig in Publikationen finden. Oft greifen Autoren auf den "billigen Ausweg" zurück, indem sie sich auf Kollegen berufen, die ähnliche Behauptungen aufgestellt haben - doch das ist keine wirkliche Evidenz.

## 2.2.2 Komplexe Fragen zur zeitgenössischen Rezeption historischer Hypothesen

Stellen Sie sich vor, Sie möchten herausfinden, welcher zeitgenössische Autor sich zu einer spezifischen These des Wissenschaftshistorikers Johannes Kepler aus dem Jahr 1603 geäußert hat. Ohne jahrelange, akribische Lektüre und Archivarbeit wäre es unmöglich, eine solche Frage zu beantworten. Ähnlich verhält es sich mit Aussagen darüber, wer die Publikation einer historischen Hypothese maßgeblich beeinflusst hat. Solche Behauptungen halte ich meist für spekulativ und unbegründet - nicht weil unseriös geforscht wurde, sondern weil der Nachweis der Evidenz extrem schwierig ist.

## 2.3 Die Bedeutung der AI für die Geisteswissenschaften

Die AI bietet uns nicht nur technische Erleichterungen im Forschungsalltag, sondern ermöglicht es uns auch, bisher nur unzureichend lösbarer Fragestellungen endlich fundiert zu bearbeiten. Dazu gehören auch Fragen nach Alternativen zu historischen Hypothesen oder nach der Nachvollziehbarkeit und Überzeugungskraft von Begründungen für Zeitgenossen.

In den kommenden Jahren wird die AI unsere wissenschaftlichen Disziplinen drastisch verändern. Daher rate ich Ihnen dringend, sich schon während des Studiums mit diesen Mitteln vertraut zu machen, um den künftigen Anforderungen gerecht zu werden.

## 2.4 Die Evolution der Mensch-Maschine-Interaktion

Die Art und Weise, wie wir mit künstlicher Intelligenz interagieren, hat sich in den letzten Jahrzehnten kontinuierlich weiterentwickelt. Vor etwa 25 Jahren revolutionierte die Erfindung des Browsers unser Informationszeitalter. Plötzlich konnten wir über Verlinkungen auf ein schnell wachsendes Netzwerk an Informationen zugreifen. Rund 15 Jahre später folgte das Smartphone, das heute aus unserem Alltag nicht mehr wegzudenken ist.

### 2.4.1 Von der Adresseingabe zur Suchanfrage

Das ursprüngliche Adressfeld zur Eingabe von Weblinks hat sich im Laufe der Zeit zu einem mächtigen Werkzeug entwickelt, mit dem wir beliebige Suchanfragen stellen können. Suchmaschinen wie Google verarbeiten unsere Eingaben und liefern uns die gewünschten Ergebnisse.

### **2.4.2 Der Durchbruch von Chat-GPT**

Mit der Einführung von Chat-GPT erleben wir gerade einen massiven Umbruch in der Interaktion zwischen Mensch und Maschine. Statt mit einem Provider zu kommunizieren, interagieren wir nun mit einem KI-Modell, das unsere Informations- und Mitteilungsbedürfnisse steuert. Dieser Paradigmenwechsel hat tiefgreifende Auswirkungen auf die Art und Weise, wie wir auf Wissen zugreifen und es verarbeiten.

### **2.4.3 Neue Schnittstellen: Sprache, Gesten und Gedanken**

Die Möglichkeiten der Mensch-Maschine-Interaktion entwickeln sich rasant weiter. Sprachbefehle, wie wir sie von Siri kennen, ermöglichen es uns, Computer per Spracheingabe zu steuern. Datenbrillen und Headsets eröffnen neue Perspektiven, indem sie uns kontextbezogene Informationen in Echtzeit liefern. Selbst Gesten und Hirnströme können als Eingabesignale genutzt werden. Wohin diese Entwicklung führt, lässt sich nur schwer vorhersagen, aber eines ist sicher: Die Zukunft der Mensch-Maschine-Interaktion verspricht spannende Möglichkeiten.

## **2.5 Die Macht der generativen AI**

Hinter all diesen faszinierenden Anwendungen steckt die sogenannte generative AI oder kurz Gen-AI. Dieser Ansatz ermöglicht es, bedeutungsvolle sprachliche Ausdrücke zu erzeugen - ein revolutionärer Schritt, den es in dieser Form zuvor nicht gab.

### **2.5.1 Von der Syntax zur Semantik**

Bisher beschränkte sich der Umgang von Computern mit unserer sprachlichen Welt auf die Verarbeitung von Zeichenketten, die bestimmte syntaktische Regeln erfüllten. Jetzt kommt jedoch die Semantik ins Spiel - die Bedeutung dieser Zeichen. Hier eröffnet sich ein völlig neues Feld für die Philosophie.

### **2.5.2 Die Bedeutung sprachlicher Ausdrücke**

Sprachliche Ausdrücke sind sinnlich wahrnehmbare Zeichen, die eine Bedeutung tragen. Im Gegensatz zu bloßen materiellen Dingen in der Welt, die keine Zeichen sind, verweisen sprachliche Ausdrücke auf etwas. Genau hier setzt die semantische Dimension an.

### 2.5.3 Philosophische Kritik an der Terminologie

Die großmäulige Propaganda der Konzerne, die schon von “Knowledge Graphen” sprachen, als von Bedeutung noch keine Rede war, sollte philosophisch hinterfragt werden. Bei näherer Betrachtung entpuppt sich dieses Kartenhaus als Unsinn - es handelt sich um einfache Graphen, nicht um “Knowledge Graphen”. Die philosophische Kritik entlarvt, was sich hinter solchen Begrifflichkeiten verbirgt und stellt die Frage, was es eigentlich heißt, von der Bedeutung einer Computeraussage zu sprechen.## Einführung in die semantische Revolution der AI

Meine sehr verehrten Damen und Herren, lassen Sie uns heute gemeinsam einen faszinierenden Blick in die aktuellsten Entwicklungen der AI-Technologie werfen. Hier geht es um nichts Geringeres als um den Kern der AI-Revolution: Die Fähigkeit, sprachliche Ausdrücke, Zeichen und Symbole mit ihrer Bedeutung zu verbinden. Wie gelingt es der AI auf einmal, diese Verknüpfung herzustellen? Und welche weitreichenden Konsequenzen ergeben sich daraus? Das sind die spannenden Fragen, denen wir uns heute widmen werden.

### 2.5.4 Von der Zeichenkettensuche zur Bedeutungsanalyse

Stellen Sie sich vor, Sie nutzen eine herkömmliche Suchmaschine wie Google. Was passiert, wenn Sie einen Suchbegriff eingeben? Die Maschine durchforstet raffiniert, aber letztlich mechanisch, riesige Datenbestände nach passenden Zeichenketten, Adressen, Wortbegriffen oder Namen. Damit lässt sich zweifellos Beachtliches erreichen, aber im Kern bleibt es eine Suche nach Zeichenfolgen.

Doch nun eröffnet sich eine völlig neue Dimension: Die Suche nach Aussagen, nach Inhalten von Ausdrücken. Das ist ein fundamentaler Unterschied. Lassen Sie uns das an einem einfachen Beispiel verdeutlichen: “Der Hund ist schwarz.” Dieser Satz, den ich gerade ausgesprochen habe, ist zunächst einmal eine Zeichenkette. Syntaktisch korrekt, aber noch kein Inhalt an sich. In der Philosophie unterscheiden wir sehr genau zwischen dem Satz und seiner Bedeutung.

### 2.5.5 Wahrheitswerte und die Welt der Aussagen

Und hier kommt der entscheidende Punkt: Sätze selbst sind weder wahr noch falsch. Sie sind sprachliche Ausdrücke, die wohlgeformt sein können, aber keine Wahrheitswerte besitzen. Wahr oder falsch sind die mit Sätzen ausgedrückten Inhalte, die wir in der Philosophie als Aussagen, Propositionen oder Statements bezeichnen.

Solange wir uns nur in der Welt der Syntax bewegen, haben wir es noch nicht einmal mit der Ebene des Wahren und Falschen zu tun. Und wenn wir nicht in der Welt des Wahren und Falschen sind, können wir auch nichts glauben. Denn wir glauben nur etwas, wenn wir von etwas sprechen, das wahr oder falsch sein kann. Erst dann können wir Überzeugungen entwickeln und etwas für richtig oder falsch halten.

Doch genau hier setzt die AI-Revolution an. Mit den neuen technischen Mitteln bewegen wir uns plötzlich in der Dimension der Aussagen. Eine völlig neue Welt tut sich auf. Aussagen sind die Träger von Wahrheitswerten. Und erst wenn wir von Aussagen mit Wahrheitswerten sprechen, kommen Begriffe wie Rechtfertigung, Kritik oder Widerlegung ins Spiel. Die gesamte erkenntnistheoretische Dimension des Wissens, des Behauptens, Findens, Kritisierens und Widerlegens setzt voraus, dass wir es mit Aussagen und ihren Wahrheitswerten zu tun haben.

### **2.5.6 Die Dimension der Aussagen eröffnet neue Möglichkeiten**

Sie sehen, welch weitreichende Konsequenzen sich daraus ergeben. Eine Suchmaschine, die nur Zeichenketten findet, lässt sich nicht sinnvoll kritisieren. Sie hat ihre Aufgabe erfüllt, wenn sie passende Strings gefunden hat. Doch sobald wir in die Dimension der Aussagen vordringen, eröffnen sich ganz neue Möglichkeiten. Plötzlich können wir Maschinen befragen, ob ihre Antworten wahr oder falsch sind. Wir können ihre Aussagen hinterfragen, rechtfertigen oder widerlegen.

Früher hätte man gesagt, dass dafür der menschliche Geist, der Verstand oder die Vernunft notwendig seien. Doch nun scheinen Maschinen in der Lage zu sein, belastbare Entscheidungen zu treffen, Aussagen zu generieren, die Konsequenzen für unser alltägliches Leben haben. Das sind faszinierende Perspektiven, die sich hier auftun und die wir in den kommenden Vorlesungen noch vertiefen werden.

Doch lassen Sie uns zunächst der Frage nachgehen, wie es der AI gelingt, in die Welt der Semantik vorzudringen. Welche Techniken und Verfahren ermöglichen diesen Quantensprung?

## **2.6 Die drei Säulen der semantischen Revolution**

Ich möchte die Revolution, von der wir hier sprechen, in drei Teilespekte gliedern - drei Säulen, wenn Sie so wollen, auf denen die semantischen Fähigkeiten der AI-Modelle beruhen.

## 2.6.1 1. Das Training mit bedeutungsähnlichen Begriffen

Die erste Säule ist das Training der AI-Modelle, bedeutungsähnliche Begriffe, Sätze und Ausdrücke zu unterscheiden. Lassen Sie uns das an einem Beispiel veranschaulichen:

- “An eagle flies silently over the large tree.”
- “A swan flies noisily over the large tree.”
- “A mouse eats happily a piece of cheese.”

Intuitiv erkennen wir sofort, dass die ersten beiden Sätze semantisch ähnlich sind, auch wenn sie sich in Details unterscheiden. Im dritten Satz hingegen geht es um etwas völlig anderes, obwohl auch hier ein Tier eine Handlung ausführt.

### 2.6.1.1 Embeddings als Grundlage der Bedeutungsanalyse

Doch wie gelingt es der AI, diese Ähnlichkeiten und Unterschiede zu erfassen? Die Antwort liegt in sogenannten Embeddings. Dabei handelt es sich um mathematische Repräsentationen, die den Verwendungszusammenhang von Wörtern in einem riesigen Textkorpus erfassen.

Durch das Training mit Billionen von Worteinheiten aus dem Internet erstellen die AI-Modelle gigantische Tabellen, die für jedes Wort festhalten, in welchem Kontext es typischerweise auftritt, welche Wörter ihm vorangehen und folgen. Durch mathematische Verfahren lassen sich diese Tabellen so komprimieren, dass am Ende eine überschaubare Zahl von Dimensionen ausreicht, um die Bedeutungsrolle jedes Wortes in einem Satz zu erfassen.

Mit Hilfe dieser Embeddings kann die AI dann beurteilen, welche Sätze semantisch ähnlich sind. Sie liefert sogar einen numerischen Wert für den Grad der Ähnlichkeit. Dabei geht es zunächst noch nicht um die eigentliche Bedeutung, sondern um die Kombinationshäufigkeit der Wörter untereinander. Aber es ist ein entscheidender Schritt auf dem Weg zur Erfassung von Bedeutungsaspekten.

## 2.6.2 2. Die Frage nach der Bedeutungsgleichheit

Die zweite Säule der semantischen Revolution ist die Fähigkeit der AI, bedeutungsgleiche Ausdrücke zu erkennen. Welche sprachlichen Ausdrücke, die sich in ihrer Syntax unterscheiden, drücken dennoch das-selbe aus, haben denselben Wahrheitswert?

### **2.6.2.1 Aktiv-Passiv-Transformation und Übersetzung**

Zwei klassische Beispiele für bedeutungsgleiche Ausdrücke sind die Aktiv-Passiv-Transformation und die Übersetzung. „Der Hund jagt die Katze“ und „Die Katze wird vom Hund gejagt“ mögen sprachlich verschieden sein, bedeuten aber dasselbe. Ebenso verhält es sich mit „Der Hund ist schwarz“ und „The dog is black“. Jedes Wort ist anders, doch die Aussage bleibt gleich.

Früher war die Computerlinguistik mit dieser Herausforderung weitgehend überfordert. Doch heute gehört die maschinelle Übersetzung zur Grundausrüstung der AI-Modelle. Und das nicht nur Wort für Wort, sondern unter Berücksichtigung komplexer grammatischer und stilistischer Anforderungen, wie es ein guter menschlicher Übersetzer tun würde.

### **2.6.2.2 Trainingsdaten aus Übersetzungsliteratur und Philosophie**

Doch wie wurde dieses erstaunliche Können erreicht? Ein Schlüssel liegt in den Trainingsdaten. Die AI-Modelle wurden mit den besten verfügbaren Übersetzungen trainiert, von den Klassikern der Weltliteratur bis hin zu philosophischen Texten.

Gerade die philosophische Literatur erwies sich als unschätzbare Quelle, denn hier finden sich präzise sprachphilosophische Reflexionen über die Inhalte von Aussagen. Was sind logische Schlussformen? Welche Regeln gelten für das semantische Schließen? All das ist in den Lehrbüchern der Logik zu finden, die nun zum Standardrepertoire der AI-Modelle gehören.

### **2.6.3 3. Das Training mit logischen Regeln**

Damit sind wir bei der dritten Säule angelangt: dem Training der AI mit den Regeln der Logik. So wie Philosophiestudierende in den Einführungsvorlesungen die Grundlagen des logischen Schließens erlernen, so haben auch die AI-Modelle diese Regeln verinnerlicht.

Ein Modus ponens gehört ebenso zum Repertoire der AI wie für angehende Philosophen. Natürlich gibt es noch Fälle, in denen die Maschinen daneben liegen. Aber die Fortschritte sind beeindruckend und eröffnen faszinierende Perspektiven.

## 2.7 Ausblick

Meine Damen und Herren, wir haben heute einen ersten Einblick in die semantische Revolution der AI gewonnen. Wir haben gesehen, wie durch Embeddings, Übersetzungstraining und logische Regeln die Grundlagen geschaffen wurden, dass Maschinen in die Welt der Bedeutungen vordringen können.

Die Konsequenzen sind weitreichend und werden uns noch lange beschäftigen. Können Maschinen wirklich Aussagen treffen, die für unser Leben relevant sind? Welche ethischen Fragen wirft das auf? Und wo liegen die Grenzen dieser Technologie?

Das sind spannende Fragen, denen wir uns in den kommenden Vorlesungen widmen werden. Ich freue mich darauf, gemeinsam mit Ihnen tiefer in diese faszinierende Materie einzutauchen und die Möglichkeiten und Herausforderungen der AI-Revolution zu erkunden.## Bedeutungsähnlichkeit und die Revolution der Künstlichen Intelligenz

Meine sehr verehrten Damen und Herren, lassen Sie uns heute einen tieferen Blick in die faszinierende Welt der Künstlichen Intelligenz werfen - eine Welt, die von bahnbrechenden Entwicklungen geprägt ist, welche die Art und Weise, wie wir mit Sprache und Bedeutung umgehen, grundlegend verändern. Im Zentrum dieser Betrachtung steht das Konzept der Bedeutungsähnlichkeit und wie es die KI-Landschaft revolutioniert hat.

### 2.7.1 Die Bedeutung von Embeddings

Embeddings, numerische Repräsentationen sprachlicher Ausdrücke, bilden das Fundament für die Zuordnung von Bedeutung in der KI. Doch es ist wichtig zu verstehen, dass sie lediglich die Vorstufe des Trainings darstellen und nicht als rigide Objekte missverstanden werden dürfen. Die wahre Bedeutung von Ausdrücken lässt sich oft nur im Kontext ihrer Verwendung beurteilen - eine Erkenntnis, die uns vor vorschnellen Schlüssen bewahrt.

### 2.7.2 Die Suche nach bedeutungsähnlichen Aussagen

Stellen Sie sich vor, Sie fragen eine KI: "Fliegt da ein Schwan über den Baum?" Was passiert nun im Hintergrund? Die KI übersetzt diesen Satz in eine numerische Repräsentation, ein Embedding in tausenden Dimensionen. Mit dieser Zahl durchsucht sie dann eine Datenbank nach Büchern, in denen ähnliche Aussagen formuliert werden - unabhängig von der Sprache oder syntaktischen Transformationen. Plötz-

zlich können wir die gesamte Literatur nach Inhalten durchforsten, nicht nur nach Zeichenabfolgen. Eine wahrhaft revolutionäre Entwicklung!

### **2.7.3 Die Erweiterung auf verschiedene Medien**

Doch damit nicht genug: Embeddings gibt es nicht nur für Texte, sondern auch für Bilder, Videos, Audio, 3D-Objekte und sogar Hologramme. Die Programme können nicht nur Texte inhaltlich verstehen, sondern auch begleitende Bilder oder Diagramme erschließen. Eine multimediale Welt der Bedeutung eröffnet sich uns.

## **2.8 Die zweite Revolution: Attention is all you need**

Der Slogan “Attention is all you need” markiert den Beginn der zweiten Revolution in der KI. In einem bahnbrechenden Artikel auf dem Preprint-Server arXiv zeigten Forscher von Google, wie man Sprache als Abfolge von Token versteht und die Aufgabe darin besteht, das nächste Wort vorherzusagen. Was zunächst trivial klingen mag, entpuppt sich als Schlüssel zu einer neuen Ära der KI.

### **2.8.1 Die Macht der Vorhersage**

Lassen Sie uns ein Beispiel betrachten: “Der Hund ist schwarz.” Was erwarten Sie als Antwort auf diese Aussage? Wahrscheinlich sind Sie genauso perplex wie ein KI-Modell, das mit einer solchen Feststellung konfrontiert wird. Die Programme haben eingebaute Attention-Mechanismen, die prognostizieren, was als nächstes kommen könnte. Bei einer schlichten Feststellung wie dieser fällt die Vorhersage schwer - ein Umstand, der zu teils kuriosen Reaktionen der KI führen kann.

### **2.8.2 Von der Frage zur Anweisung**

Die Nutzung von KI hat sich von der reinen Frage-Antwort-Interaktion hin zu Anweisungen und Instruktionen verschoben. Die Modelle wurden entsprechend umtrainiert und zu Akteuren, die Instruktionen ausführen. Der Attention-Mechanismus ermöglicht es, plausible Textfolgen als Antwort zu generieren, abhängig von der Art der Eingabe - sei es eine Frage, eine Anweisung oder eine Aussage, die eine bestimmte Reaktion hervorruft.

## 2.9 Die Komposition von Instruktionen und Inhalten

Die gegenwärtigen KI-Modelle bestehen im Wesentlichen aus der Komposition eines Vordersatzes mit Instruktionen und vielem mehr, sodass die Ausgabe im Idealfall eindeutig konstruiert werden kann. Nehmen wir das Beispiel “Übersetze den Satz ‘Der Hund ist schwarz’”. Das Programm reformuliert intern die Eingabe in eine explizite Wiedergabe des Inhalts, um alle impliziten Annahmen offenzulegen. So wird sichergestellt, dass die Übersetzung korrekt erfolgt, unabhängig von sprachlichen Nuancen oder Mehrdeutigkeiten.

Meine Damen und Herren, wir stehen an der Schwelle zu einer neuen Ära der Künstlichen Intelligenz, in der Bedeutung und Kontext eine zentrale Rolle spielen. Die Entwicklungen im Bereich der Embeddings und des Attention-Mechanismus haben die Art und Weise, wie wir mit Sprache und Wissen umgehen, grundlegend verändert. Lassen Sie uns gemeinsam diese faszinierende Reise fortsetzen und die Möglichkeiten erkunden, die sich uns eröffnen. Die Zukunft der KI ist wahrlich aufregend!## Textgenerierung und Kontext

Zunächst möchte ich Ihnen näherbringen, wie die Textgenerierung in den gegenwärtigen AI-Modellen funktioniert. Ein entscheidender Aspekt ist dabei der Kontext. Stellen Sie sich vor, ich gebe in das Programm lediglich den Satz “Der Hund ist schwarz.” ein, ohne jeglichen weiteren Kontext. Was passiert dann? Das Programm beginnt eigenständig, weitere Informationen zu generieren. Es könnte beispielsweise über schwarze Labradore schreiben und allerlei zusätzliche Kontextinformationen hinzufügen.

Genau hier liegt das Problem der sogenannten “Halluzination”. Da keine Beschränkungen hinsichtlich des Inhalts oder der sachlichen Prüfung vorgegeben sind, kann das Programm frei assoziieren und scheinbar sinnvolle Sätze generieren, die jedoch nicht unbedingt der Wahrheit entsprechen.

### 2.9.1 Sprachkompetenz vs. Sachkompetenz

Es ist wichtig zu verstehen, dass die Modelle, um die es hier geht, im Grunde nur eines beherrschen: die Übersetzung von sprachlichem Ausdruck in ihre Bedeutung. Sie verfügen über eine ausgeprägte Sprachkompetenz, aber keinerlei Sachkompetenz. Es mag zwar suggeriert werden, aber in Wirklichkeit existiert in diesen Programmen kein Mechanismus, der prüft, ob das, was als scheinbar sinnvoller Satz generiert wird, auch tatsächlich sachlich wahr ist.

Wir stehen also vor einer Revolution, bei der wir es nicht mehr nur mit Sätzen zu tun haben, sondern mit Aussagen, die durch diese Sätze ausgedrückt werden. Damit eröffnet sich die Dimension der Wahrheit, der Rechtfertigung und der Kritik. Doch die aktuellen AI-Programme lösen diese Frage nicht ein. Sie

prüfen nicht die sachliche Korrektheit, führen keine Evidenz an und kritisieren auch nicht. Das ist schlichtweg nicht Teil der Programme.

### **2.9.2 Gefahren und Grenzen von Chat-GPT**

Angesichts dieser Tatsachen möchte ich Ihnen dringend davon abraten, Hausarbeiten mit Chat-GPT zu schreiben. Die Wahrscheinlichkeit, dass die generierten Inhalte falsch sind, ist überwältigend hoch. Sie werden immer auffliegen, denn Sie selbst sind nicht in der Lage zu prüfen, ob das, was das Programm ausgibt, tatsächlich wahr ist.

Das Tückische dabei ist, dass die Programme perfekt darin sind, die Inhalte sinnvoll erscheinen zu lassen. Lassen Sie mich ein Beispiel geben: Ich stellte einmal eine anspruchsvolle historische Frage zum Publikationsverhalten von Leonhard Euler im Jahr 1756. Man würde erwarten, dass das Programm bei so spezifischen historischen Informationen zugibt, keine Antwort zu haben. Stattdessen kam eine Literaturangabe, die auf den ersten Blick perfekt aussah. Sie passte zum Autor und zu der Publikationsreihe, in der er normalerweise veröffentlichte. Sogar die Bandzahl stimmte. Doch der Titel war völlig erfunden - diese Publikation hat es nie gegeben! Selbst ich als Experte habe nicht sofort erkannt, dass es sich um eine Fälschung handelte, so perfekt war die Formatierung. Hätten Sie diese Angabe in eine Arbeit kopiert und wären kein Experte auf diesem Gebiet, hätten Sie den Schwindel nicht bemerkt.

## **2.10 Erweiterung der AI-Modelle**

### **2.10.1 Sachliche Korrektheit und Wahrheit**

Um dieses Problem anzugehen und die sachliche Korrektheit der generierten Inhalte zu gewährleisten, müssen wir uns fragen: Was fehlt den aktuellen AI-Modellen und was muss hinzugefügt werden, damit sie nicht nur Sprachkompetenz, sondern auch das Wissen der Welt besitzen?

Als Wissenschaftler müssen wir Wege finden, die Inhalte zu prüfen, zu validieren und sicherzustellen, dass sie der Wahrheit entsprechen. Stellen Sie sich vor, Sie würden selbst eine Quelle wie Eulers Publikation überprüfen wollen. Sie würden glaubwürdige Referenzen konsultieren, vielleicht Eulers Opera Omnia durchsuchen oder sogar in eine Bibliothek gehen, um die Publikation physisch in die Hand zu nehmen. Das ist das normale Vorgehen in der Gelehrtenwelt.

Doch wie könnte dies in einer zukünftigen Welt der AI aussehen? Es ist klar, dass die aktuellen Modelle dafür nicht ausreichen. Es genügt nicht, dass der Output syntaktisch wohlgeformt und plausibel

erscheint. Es fehlen entscheidende Elemente, um sachliche Korrektheit herzustellen.

### **2.10.2 Korrespondenztheorie der Wahrheit**

Eine mögliche Antwort liefert die Korrespondenztheorie der Wahrheit. Dabei geht man davon aus, dass der sprachliche Ausdruck sinnvollerweise der sachlichen Struktur in der Welt, auf die er sich bezieht, entsprechen sollte. Stimmt diese Übereinstimmung, ist die Aussage wahr, ansonsten ist sie falsch.

Doch um dies in den AI-Modellen umzusetzen, müssen zusätzliche methodische Schritte unternommen werden. Die Modelle müssen Zugriff auf das haben, worauf die Sprache eine Korrespondenzbeziehung haben sollte. Das ist die große Herausforderung, an der wir arbeiten müssen.

## **2.11 Sprachentwicklung und Bedeutungsverschiebungen**

Ein weiterer Aspekt, den es zu berücksichtigen gilt, ist die Tatsache, dass Sprache nicht statisch ist, sondern sich im Laufe der Zeit verändert. Auch diese historisch gewachsenen, kontextuell bedingten Verschiebungen von Sprache und Sprachverständnis müssen die AI-Modelle abbilden können.

Wie weit man Sprachmodelle darauf trainieren kann, ist noch Gegenstand intensiver Forschung. Es gibt erste Untersuchungen mit ausgesuchten Teilbegriffen, aber insgesamt steckt dieses Feld noch in den Kinderschuhen. Hier ist noch enorm viel Forschungsarbeit zu leisten.

### **2.11.1 Fehltraining und Sprachmarotten**

Ein Risiko besteht darin, dass sich in den AI-Modellen mehr oder weniger zufällige Sprachmarotten bilden, die quasi neu entstehen und nichts mit dem zu tun haben, was zuvor von Menschen produziert wurde. Ein Beispiel dafür sind die Open-AI-Modelle, die in einer bestimmten Trainingsphase offenbar mit Literatur trainiert wurden, die sich nicht auf sachliche kausale Relationen fokussierte, sondern auf die Überzeugungen von Personen darüber, was die Ursache von etwas ist.

Das führte dazu, dass diese Modelle nicht in der Lage waren, die üblichen Regeln des kausalen Schließens anzuwenden. Stattdessen modellierten sie letztlich, wie Personen etwas in kausaler Hinsicht über die Welt glauben. Auf die Frage "Der Hund ist schwarz." kam dann etwa die Antwort "Person B glaubt, er könnte aber auch braun sein." - obwohl danach gar nicht gefragt wurde.

Dieses Beispiel zeigt, wie entscheidend die Kontextkonstruktion bei den AI-Modellen ist. Wir machen momentan die Erfahrung, wie sich die Modelle verhalten, und es geht oft noch deutlich daneben. Wie man dies konzeptuell in den Griff bekommt, ist alles andere als klar. Aber es gibt Wege, die ich Ihnen im Laufe des Semesters aufzeigen werde.

### **2.11.2 Reichhaltige Kontextkonstruktion**

Der Schlüssel liegt darin, die Eingabetexte im sogenannten Kontext informativer und reichhaltiger zu gestalten. Dann erhält man auch entsprechend hochwertige Antworten. Wenn Sie beispielsweise merken, dass das Programm nicht nach Sachfragen, sondern nach Überzeugungen von Personen antwortet, müssen Sie explizit machen, dass Sie keine Antworten basierend auf Personenüberzeugungen wünschen. In den meisten Fällen reicht das aus, um solche Fehler zu korrigieren.

Allerdings können die Modelle manchmal sehr hartnäckig sein. Dann hilft nur noch, das Fenster rauszuschmeißen, wie man so schön sagt. Aber das sind Erfahrungswerte, die wir nach und nach sammeln.

## **2.12 Anwendungsbeispiele und Potenziale**

Lassen Sie mich zum Abschluss noch ein paar weitere Anwendungsbeispiele und Potenziale von AI-Modellen aufzeigen.

### **2.12.1 Übersetzungen als Motor des Trainings**

Übersetzungen waren nicht nur ein kulturelles Plus, sondern der eigentliche Motor des Trainings von Bedeutungsgleichheit. Die Programme sind mittlerweile in der Lage, beliebige Sätze zu übersetzen, selbst wenn die übersetzte Formulierung nirgendwo in der Literatur zu finden ist.

Nehmen wir an, ein anspruchsvolles deutsches Werk wie Goethes Faust oder ein Roman von Thomas Mann soll in eine Sprache übersetzt werden, in der es noch keine Übersetzung gibt. Die AI-Modelle können das leisten. Ob die Übersetzung dann in jeder Hinsicht perfekt ist, darüber kann man diskutieren. Aber sie werden einen Vorschlag liefern.

## 2.12.2 Zusammenfassungen und Frage-Antwort-Systeme

Ein weiteres beeindruckendes Anwendungsfeld sind Zusammenfassungen. Mittlerweile ist es möglich, ganze Bücher in das Programm einzugeben und eine Zusammenfassung für jedes Kapitel in einem Absatz zu erhalten. Die Ergebnisse sind relativ solide und belastbar.

Auch Frage-Antwort-Systeme wie die Chatbots haben ein enormes Potenzial. Hier kommt ein Aspekt zum Tragen, der häufig übersehen wird: die semantische Korrektur.

In unserer Kommunikation findet oft eine Dimension des Austauschs statt, bei der es nicht um Sachinformationen geht, sondern um die Klärung von Bedeutungen. Wir fragen "Was meinst du damit?" oder "Meinst du gerade dieses?", um sicherzustellen, dass wir die Aussage des Gegenübers richtig verstanden haben.

Genau diese Interaktionen der Bedeutungsklärung und Kontextkorrektur sind der Clou von Systemen wie Chat-GPT. Wenn Sie eine Frage stellen, beispielsweise "Wann lebte Leonhard Euler?", weiß das Programm zunächst nicht, welchen Leonhard Euler Sie meinen - den berühmten Mathematiker oder vielleicht Ihren Nachbarn, der zufällig denselben Namen trägt und eine Pommesbude betreibt.

Durch den anschließenden Dialog, in dem Sie klarstellen, dass Sie nicht den Mathematiker, sondern den Pommesbuden-Besitzer meinen, wird der Kontext der ursprünglichen Annahme korrigiert und eine verbesserte Antwort generiert.

Das bedeutet: Indem Sie chatten, tragen Sie aktiv zur künstlichen Intelligenz der Gesamtantwort bei. Auch wenn es Ihnen vielleicht nicht bewusst ist - durch das dialogische Interagieren mit dem Programm werden Sie zu einem essenziellen Teilnehmer am Entstehungsprozess der Antwort.

Diese Erkenntnis ist von großer Bedeutung und wird bis heute in der technischen Umsetzung sinnvoll genutzt und gepflegt. Und genau hier liegt meiner Meinung nach einer der spannendsten Aspekte dieser Technologie, den es in Zukunft weiter zu erforschen und zu optimieren gilt.

# **3 Charakter von LLMs**

## **3.1 Vorlesung Philosophie der AI: Generative Modelle, Large Language Models und Character-Konfiguration**

Willkommen zurück zur dritten Vorlesung der Philosophie der AI! Bevor wir tiefer in die faszinierende Welt der generativen Modelle eintauchen, lassen Sie mich einige organisatorische Aspekte ansprechen. Ich möchte Ihnen versichern, dass trotz der verwirrenden Ablehnungsbescheide des Agnes-Zulassungssystems jeder immatrikulierte Student, auch ÜWP, zu dieser Vorlesung zugelassen ist, solange wir Platz in diesem Saal haben. Die Philosophische Fakultät und ich persönlich garantieren Ihnen dies. Es ist lediglich wichtig sicherzustellen, dass Ihre Studienleistungen korrekt in das Prüfungssystem Ihres Hauptfaches eingetragen werden. Bei Fragen oder Bedenken wenden Sie sich bitte an das zuständige Prüfungsbüro oder an Frau Krause vom Sekretariat der Philosophie. Wir werden gemeinsam sicherstellen, dass Ihre Leistungen entsprechend dokumentiert werden.

Später in der Vorlesung werde ich Ihnen außerdem mögliche Projektarbeiten vorstellen, die Sie im Rahmen eines Gesamtforschungsvorhabens in Zusammenarbeit mit wissenschaftlichen Akademien und der Stiftung Deutscher Klassik in Weimar absolvieren können. Je nach Ergebnissen dieser Übungen überlegen wir, die Resultate am Ende des Semesters öffentlichkeitswirksam zu präsentieren. Ich freue mich darauf, Ihnen die Details in Kürze unterbreiten zu können.

## **3.2 Die Revolution der generativen AI-Modelle**

In den letzten beiden Stunden haben wir uns bereits intensiv mit den verschiedenen Modellen der AI oder KI auseinandergesetzt. Ein Begriff, der sich zunehmend etabliert, ist der der generativen AI-Modelle. Diese Modelle zeichnen sich dadurch aus, dass sie in der Lage sind, abhängig von einem gegebenen Input, Texte, Bilder, Videos oder Audiodaten zu erzeugen - sie generieren etwas Neues.

### **3.2.1 Large Language Models als Kern der generativen AI**

Im Kern dieser generativen AI stehen die sogenannten Large Language Models (LLM). Wie der Name schon sagt, handelt es sich hierbei um umfangreiche Sprachmodelle, die auch dann zentral bleiben, wenn es um die Verarbeitung und Interpretation von Bildern geht. Die Ebene des Sprachverständens und -verarbeitens ist fundamental für alle Modelle der künstlichen Intelligenz, mit denen wir es hier zu tun haben.

### **3.2.2 Die Explosion der verfügbaren Modelle**

Derzeit gibt es etwa 100 verschiedene Vorschläge für solche Modelle, von denen einige nur über Lizenzen und Zugriffsbarrieren nutzbar sind, während die Mehrzahl bereits Open Access zur Verfügung steht. Die Anzahl der angebotenen Modelle explodiert förmlich, wobei jedes Modell seine eigenen spezifischen Kompetenzen und Fähigkeiten aufweist.

## **3.3 Die Funktionsweise der generativen AI-Modelle**

### **3.3.1 Semantische Ähnlichkeit und Transformation**

Die derzeitige Generation der Modelle arbeitet im Wesentlichen mit zwei revolutionären Komponenten:

1. Semantische Ähnlichkeit: Die Modelle sind in der Lage, Bedeutungsgleichheiten oder -ähnlichkeiten zu identifizieren, anstatt nur nach exakten Stichwörtern zu suchen.
2. Transformation: Basierend auf diesen semantischen Ähnlichkeiten können die Modelle bei einem gegebenen Input einen passenden Output generieren.

Die Kombination dieser beiden Aspekte ist extrem weitreichend, da sie eine Verallgemeinerung der Bedeutung von Textinhalten und eine Transformation dieser Regeln ermöglicht. Ähnlich wie wir Menschen allgemeine Regeln aufstellen können, sind diese Modelle in der Lage, verallgemeinerte Regeln zu erkennen und anzuwenden.

### **3.3.2 Character - Die Formung des künstlichen Charakters**

Zusätzlich zu den Sprachkompetenzen kommt nun ein dritter Aspekt ins Spiel, der die Philosophie auf den Plan ruft: der sogenannte "Character". Die generativen AI-Modelle verhalten sich in der Kommunikation

fast so, als würde man mit einer menschlichen Person interagieren. Durch die Beherrschung der semantischen Verallgemeinerung und der Regeltransformation können wir die Art und Weise, wie diese Modelle Regeln erstellen, modifizieren und sie so charakterlich formen.

### **3.3.2.1 Stilistische Aspekte des Characters**

Diese Charakterformung kann sehr weit reichen und umfasst zunächst oberflächliche, stilistische Aspekte wie:

- Sprache der Antworten
- Schreibstil (z.B. im Stil von Ernest Hemingway)
- Datenausgabe (knapper, schematisiert, in bestimmten Formaten)
- Literarische Stile (z.B. griechische Hexameter, Stil eines Homer)

### **3.3.2.2 Philosophische Aspekte des Characters**

Noch interessanter wird es bei den philosophischen Aspekten des Characters, also den formalen inhaltlichen Regeln des Nachdenkens, Resonierens und Formulierens von Arbeitsverfahren und Denkprozessen. Hier geht es darum, welche Regeln diese Modelle befolgen sollen, um die gegebenen Instruktionen zu erfüllen. Dieser Aspekt ist in der derzeitigen Entwicklung noch unterbelichtet, obwohl alle Entwickler wissen, dass er berücksichtigt werden muss.

### **3.3.3 Metaregeln und kausales Schließen**

Ein wichtiger Teilbereich der Charakterformung sind die Metaregeln, insbesondere im Bereich des kausalen Schließens. Für viele wissenschaftliche und nicht-wissenschaftliche Bereiche, wie etwa die Medizin, ist dies von großer Bedeutung. Fragen der Diagnostik, der Vorstellungen über Krankheiten und Krankheitsverläufe erfordern kausales Schließen. Bisher sind diese Regeln in den Modellen nicht systematisch vorhanden, sondern werden lediglich durch das Training anhand von Publikationen antrainiert. Die Ableitung allgemeiner Metaregeln zum korrekten Schließen und zu wissenschaftlichen Verfahren ist eine der großen Herausforderungen für die Zukunft.

### **3.3.4 Historisches Schließen**

Ein weiterer interessanter Bereich, gerade für die historischen Wissenschaften, ist das historische Schließen. Wenn es darum geht, historische Aussagen über Biografien bekannter Persönlichkeiten zu treffen, wer was erlebt und geprägt hat, sind spezifische Regeln gefragt. Auch diese Regeln müssen den Programmen erst noch beigebracht werden. Bisher haben sie nur anhand von Beispielen gelernt, einen kleinen Bereich anzuwenden, der jedoch in seinen Qualitäten limitiert ist und formale Trainingszusatzfunktionen erfordert. Ich bin zuversichtlich, dass diese Probleme innerhalb der nächsten zwei Jahre gelöst sein werden.

### **3.3.5 Die Bedeutung des Kontexts**

Neben den Regeln spielt auch der sogenannte Kontext eine entscheidende Rolle für den Input der Transformation von generativen Modellen. Der Kontext umfasst alle sprachlich ausgedrückten Zusatzinformationen, die das Programm benötigt, um zusätzlich zu einer bestimmten Instruktion einen entsprechenden Output zu generieren. Je größer und präziser dieser Kontext ist, desto besser kann die eigentliche Aufgabe inhaltlich korrekt verstanden und gelöst werden.

#### **3.3.5.1 Technische Herausforderungen des Kontexts**

Eine der interessantesten technischen Herausforderungen ist es, die Größe des Kontexts maximal zu gestalten, ohne dabei die Größe des Modells exponentiell wachsen zu lassen. Denn mit einem zu großen Modell steigen auch die Anforderungen an Hardware, Software und Stromverbrauch, was die praktische Anwendbarkeit einschränkt. Es gilt also, die richtige Balance zwischen Kontextgröße und Modellgröße zu finden, um optimale Ergebnisse zu erzielen, ohne die Bearbeitungsdauer und die Ressourcen übermäßig zu beanspruchen.## Kontextvergrößerung und Sachkompetenz bei AI-Modellen

In den letzten Monaten hat sich in der Welt der künstlichen Intelligenz viel getan. Täglich verfolge ich die Entwicklungen und bin fasziniert von den Fortschritten, aber auch besorgt über die Herausforderungen, die sich dabei auftun. Ein Stichwort, das mir besonders im Gedächtnis geblieben ist, lautet "RAG" - zusätzliche Ressourcen als Extra-Input für den Kontext. Die Idee dahinter ist, den AI-Modellen mehr Informationen zur Verfügung zu stellen, um ihre Sachkompetenz zu erweitern. Doch obwohl in den letzten fünf Monaten intensiv daran geforscht wurde, bleiben die Ergebnisse meiner Meinung nach oberflächlich und unzureichend.

Die Sachkompetenz ist eine der interessantesten zusätzlichen Anforderungen an AI-Modelle, doch aus prinzipiellen Gründen verfügen sie derzeit nicht darüber. Stattdessen kaschieren sie dieses Defizit oft geschickt. Als Warnung an alle, die Informationen von AI-Modellen nutzen: Auch wenn die Antworten überzeugend und plausibel klingen, unterliegen sie keinerlei Sachprüfung. Die Wahrscheinlichkeit, dass sie falsch sind, ist sehr hoch.

### **3.3.6 AGI - Ein umstrittenes Konzept**

Immer wieder tauchen in der Debatte um künstliche Intelligenz modische Schlagworte auf, die ebenso schnell wieder verschwinden. Ein Beispiel dafür ist die “Artificial General Intelligence” (AGI). Erst letztes Jahr erschien in der New York Times eine Stellungnahme von Kollegen, die argumentierten, warum AI prinzipiell nicht intelligent sein kann. Ihr Hauptargument: Es handle sich lediglich um probabilistische Rechnungen, die auf Wahrscheinlichkeiten basieren. Doch dieses Argument lässt sich auch auf das menschliche Gehirn übertragen - letztlich sind auch dort elektrische Impulse zwischen Neuronen für unsere kognitiven Leistungen verantwortlich. Selbst wenn diese Impulse deterministisch wären, wäre das kein Gegenargument dagegen, dass die daraus resultierenden Leistungen dem entsprechen, was wir als intelligentes Handeln und Denken bezeichnen.

Die optimistische Gegenreaktion auf solche Kritik lautet oft, dass die Entwicklung schneller voranschreiten wird als erwartet - so wie beim Schachspiel, wo Computer mittlerweile menschliche Großmeister übertreffen. Manche prophezeien, dass es bald Modelle geben wird, die das gesamte Spektrum der menschlichen kognitiven Leistungsfähigkeit überholen werden. Doch angesichts der enormen Dynamik in diesem Bereich halte ich es für unseriös, weitreichende Prognosen über die nächsten Monate hinaus abzugeben.

Ob es jemals eine Computerleistung geben wird, die alle kognitiven Leistungsbereiche des Menschen übersteigt, halte ich für eine müßige Frage. Diese Debatte gab es schon vor 30 Jahren, als klar war, dass Maschinen beim Textverständnis nicht annähernd mit Menschen mithalten konnten. Gleichzeitig konnten Computer aber bereits meisterhaft numerische Mathematik betreiben und beispielsweise Differenzialgleichungen lösen - eine Leistung, zu der kein Mensch in der Lage wäre. In vielen Bereichen der technisch-mathematischen Informatik bringen maschinelle Verfahren heute ein so hohes Problemlösungsvermögen mit, dass kein individueller Mensch mehr dagegen antreten kann. Einzelne Sektoren werden also zweifellos durch maschinelle Verfahren wesentlich kompetenter und sicherer gelöst als durch menschliche Akteure.

### 3.3.7 Hermeneutik als Herausforderung für AI

Ein Bereich, der für die Geisteswissenschaften von besonderer Bedeutung ist, ist die Interpretation von Texten. Dabei geht es darum, den Inhalt kritisch zu hinterfragen und zu verstehen - eine Leistung, die bisher dem Menschen vorbehalten war. Ziel ist es, zu einem Textverständnis zu gelangen, das nicht nur auf der Lektüre von Trainingsdatenbeständen beruht, sondern auf echten Interpretationsleistungen. Dazu müssen hermeneutische Verfahren, wie sie jeder Geisteswissenschaftler bei der Lektüre seiner Quellen anwendet, auch im Computer-Kontext umgesetzt werden. Ich habe keinen Zweifel daran, dass dies eines Tages möglich sein wird.

Doch was nützt es, solche Leistungen zur AGI hinzuzuzählen oder nicht? Manche Modelle können etwas, andere nicht - das stellen wir gerade in der Entwicklung der generativen AI-Modelle fest. Aufgrund ihrer Trainingsgeschichte haben viele Modelle beispielsweise ein ansehnliches Verständnis von Latein, obwohl der praktische Nutzen dafür gering ist. Doch diese Kompetenzen könnten schon bald wieder verschwinden, wenn die Modelle optimiert werden, um auch auf Smartphones zu laufen. Diese Optimierung bedeutet eine Reduzierung der Kompetenzen auf das Nötigste - das Gegenteil einer Entwicklung hin zu einer allgemeinen Kompetenz. Stattdessen erwarte ich eine zunehmende Spezialisierung der Modelle auf bestimmte Aufgaben wie Rechnen, Sprachverständnis oder diagnostisches Denken.

### 3.3.8 Kontextvergrößerung als Schlüssel zum Verständnis

Wenn wir über den Kontext sprechen, meinen wir ganz schlicht die Anzahl der Token (Wörter und Satzzeichen), die ein Modell berücksichtigen kann, um den Sinn einer Anfrage zu verstehen. Vor einem halben Jahr lag diese Zahl bei etwa 1.000 - das entspricht ungefähr drei Seiten Text. Alles darüber hinaus wurde nicht berücksichtigt. Wenn man also Informationen aus längeren Texten wie Enzyklopädie-Einträgen benötigte, war es unmöglich, diese vollständig in den Kontext der Modelle einzubringen. Irgendwo wurde notwendigerweise abgeschnitten und Informationen gingen verloren.

In den letzten sechs Monaten wurde daher intensiv daran gearbeitet, den Kontext zu vergrößern. Die Standardmodelle, die ich für die Illustration in dieser Vorlesung nutze, stammen aus dem Bereich "Cloth" (geschrieben wie das englische Wort für Tuch) und haben mittlerweile einen Kontext von 200.000 Wörtern. Das ist schon eine beachtliche Menge, in der sich viele Informationen unterbringen lassen.

Doch auch hier gibt es Vortäuscher, die einen großen Kontext suggerieren, ihn aber faktisch nicht nutzen. Man muss immer kritisch hinterfragen, ob die angegebene Kontextgröße auch wirklich gleichermaßen bei der Suche nach einer Antwort berücksichtigt wird.

### 3.3.8.1 Der Heunadeltest

Ein praktischer Test dafür ist der sogenannte Heunadeltest. Die Idee ist folgende: In einem beliebigen Text, beispielsweise Goethes gesammelten Werken, fügt ein Nutzer an einer Stelle einen selbst gewählten Satz oder eine Formulierung ein. Das könnte etwas sein, das Goethe nie geschrieben hätte, wie "Trump ist blöd". Die Aufgabe für das Modell besteht dann darin, genau diese Feststellung - nicht wortgleich, sondern inhaltlich - wiederzufinden. Es geht also darum, die Nadel im Heuhaufen zu finden.

Man weiß nur, dass Goethe irgendwo in seinen gesammelten Werken eine Äußerung zu Trump getätigt hat, kennt aber weder den genauen Wortlaut noch die Stelle. Vielleicht wird Trump nicht einmal namentlich erwähnt, sondern nur als "der Präsident, der 2018 im Amt war" umschrieben. Diese Nadel im Heuhaufen zu finden, ist eine anspruchsvolle Aufgabe. Es reicht nicht, einen großen Textbestand zu beherrschen - man muss nach etwas suchen, dessen Bedeutung man kennt, aber dessen genauen Wortlaut nicht.

An solchen Tests lässt sich gut erkennen, ob die verwendeten Modelle tatsächlich die Größe des Kontextes haben, die nötig ist, um einen gesamten Textbestand zu berücksichtigen. Es wäre nicht erlaubt, den Gesamttext in praktikable Teile zu unterteilen und nur in diesen zu suchen. Wenn, dann muss die Suche in Toto erfolgen. Und Goethes gesammelte Werke umfassen definitiv mehr als 200.000 Wörter. Das sprengt das Leistungsvermögen der meisten, wenn nicht aller mir bekannten Modelle.

### 3.3.9 Ausblick

Solche spezifischen Aufgaben sind meiner Meinung nach eine wesentlich bessere Beurteilung der Leistungsfähigkeit von AI-Modellen als generelle Kriterien wie AGI. Ein Katalog von Herausforderungen, die ein Modell meistern muss, um eine bestimmte Hürde zu überschreiten - das scheint mir der richtige Weg zu sein, um die Entwicklung voranzutreiben und zu bewerten.

In der nächsten Vorlesung werden wir uns genauer mit den sprachlichen Ausdrücken beschäftigen, die als Auslöser für bestimmte Reaktionen der Modelle dienen. Diese Instruktionen spielen eine entscheidende Rolle für das Verständnis und die Fähigkeiten der AI. Ich freue mich darauf, dieses faszinierende Thema mit Ihnen zu erkunden.## Die Bedeutung von Instruktionen für AI-Modelle

In der Welt der künstlichen Intelligenz spielen Instruktionen eine entscheidende Rolle. Sie sind das Herzstück der Interaktion zwischen Mensch und Maschine, denn sie geben den AI-Modellen die nötigen Anweisungen, um eine Aufgabe adäquat zu lösen. Doch nicht jede Aussage eignet sich als Instruktion. Eine simple Feststellung wie "Der Hund ist schwarz" suggeriert nichts, legt nichts nahe und fordert nicht dazu auf, etwas zu tun. Sie ist zu allgemein und vage, als dass man sie sachlich beurteilen könnte.

Die meisten AI-Modelle sind darauf trainiert, auf jede Anfrage eine Antwort zu generieren, selbst wenn die Instruktion unklar oder nicht-kommunikativ ist. Hier zeigen sich die Unterschiede zwischen den verschiedenen Modellen in der Art und Weise, wie sie mit solchen Situationen umgehen.

### **3.3.10 Von der Query zur Instruktion**

Vor einem Jahr waren Queries, ähnlich wie Google-Anfragen, noch sehr populär. Doch heute haben Instruktionen diese abgelöst und einen allgemeineren Aufgabenbereich eröffnet. Instruktionen sind derzeit das wichtigste Phänomen bei der Übergabe von sprachlich artikulierten Aufträgen an AI-Modelle.

Im Kern geht es darum, dass die Modelle in der Lage sein müssen, Instruktionen zu verstehen und auszuführen. Philosophisch gesehen handelt es sich um Handlungsanweisungen, die auf verschiedenste Anwendungsbereiche abzielen und die Modelle dazu anleiten, entsprechende Lösungen zu generieren.

## **3.4 Die Schlüsselemente der Revolution: Semantische Ähnlichkeit und regelhafte Textgenerierung**

Die Revolution in der Ausführung von Instruktionen besteht im Wesentlichen aus zwei Komponenten: der semantischen Ähnlichkeit und der Kombination mit regelhafter Textgenerierung. Doch wie wichtig der Kontext dabei ist, möchte ich Ihnen anhand einiger Beispiele verdeutlichen.

### **3.4.1 Wer war Johann Wolfgang Goethe? - Eine typische Google-Frage**

Beginnen wir mit einer Frage, die man normalerweise in eine Suchmaschine eingeben würde: "Wer war Johann Wolfgang Goethe?" Wenn wir diese Frage in das AI-Modell eingeben, erwarten wir eine Antwort, die sachlich detailliert und informativ ist. Und genau das liefert das Modell auch.

Aber woher stammen diese Informationen? Die Antwort ist einfach: aus den Trainingsdaten. Alle großen AI-Modelle wurden auf der gesamten Wikipedia, auf Millionen von wissenschaftlichen Publikationen und auf Übersetzungskorpora, einschließlich deutsch-englischer Werke, trainiert.

#### **3.4.1.1 Die Herausforderung der epistemischen Qualität**

Doch obwohl die Antwort des Modells auf den ersten Blick sehr fundiert wirkt, fehlt etwas Entscheidendes: die epistemische Qualität. Die Informationen wurden zwar verarbeitet, aber nicht systematisch auf ihre

Korrektheit geprüft. Die Modelle haben keinerlei Mittel, Falschinformationen zu erkennen.

Das ist eine Herausforderung, an der wir intensiv arbeiten. Denn obwohl die Sprachkompetenz der Modelle beeindruckend ist, mangelt es noch an echter Sachkompetenz.

### **3.4.2 Die Grenzen der Aktualität**

Ein weiteres Problem ist die Aktualität der Daten. Meistens hören die Daten, auf denen die Modelle trainiert wurden, ab einem gewissen Datum auf. Auch wenn sich die Entwickler bemühen, die Modelle zu aktualisieren, heißt das nicht, dass wirklich alle Informationen berücksichtigt und abgewogen wurden.

#### **3.4.2.1 Widersprüchliche Informationen - eine logische Herausforderung**

Viele Informationen sind widersprüchlich und damit muss man umgehen. Das ist philosophisch extrem interessant, denn aus einem Widerspruch kann man logisch gesehen alles schlussfolgern. Logisches Schließen allein löst dieses Problem nicht. Man muss präferieren.

### **3.4.3 Interne Präferenzordnungen und Regeln**

Im Hintergrund arbeiten die Modelle mit langen Listen von Alternativen zu verschiedensten Bereichen. Es gibt Präferenzordnungen, die den Modellen beigebracht wurden, um mit alternativen Antworten umzugehen. Dazu gehören auch intern trainierte allgemeine Präferenzregeln.

## **3.5 Die Qualität der Internetressourcen reicht nicht aus**

Kommen wir zurück zur Frage der Qualität von Informationen. Es gibt unterschiedliche Meinungen darüber, ob die Technologie allein ausreicht, um Informationen zu verifizieren. Ich bin da anderer Ansicht.

#### **3.5.1 Die Notwendigkeit seriöser Quellen**

Für viele entscheidende Fragen, gerade im historischen Bereich, braucht man faktisches Wissen in Details, das man sehr umfangreich suchen muss. Das Internet allein ist kein Qualitätsauszeichnungsmerk-

mal. Deswegen werden Internetquellen an Universitäten auch nicht als seriöse wissenschaftliche Quellen akzeptiert.

Man muss seine Nachweise nach den Regeln der Kunst sachlich korrekt und gerechtfertigt ausweisen. Ein simpler Internetverweis reicht da nicht. Das liegt nicht an Konkurrenzdenken, sondern an der oft mangelhaften Qualität der Informationen im Internet.

## **3.6 Die Herausforderung: Wahrheit und Wissen**

Es geht letztlich darum, Informationen zu finden, die nach bestem Wissen und Gewissen sachlich korrekt und plausibel wahr sind. Dabei geht es nicht um unumstößliche Fehlerfreiheit, sondern um Wissen, das Wahrheit impliziert. Dieses Wissen zu erlangen, ist ein Wert an sich.

### **3.6.1 Der wissenschaftliche Prozess**

Die historische Entwicklung der Wissenschaft hat über Jahrtausende Verfahren herausgearbeitet, wie man in einer großen Gruppe von Spezialisten ein kritisches Potenzial entwickelt, um maximal plausible, korrekte Antworten zu finden. Dieser Prozess ist reguliert und nicht trivial. Es geht nicht um simple Meinungsumfragen oder Mehrheitsentscheidungen.

### **3.6.2 Die offene Frage: Der Umgang mit alternativen Lösungen**

Wie geht man aber nun mit einer Mehrzahl an gerechtfertigten alternativen Lösungsvorschlägen um? Das ist eine Frage, die ich für eine spätere Vorlesung offen lassen möchte. Kein aktuelles AI-Modell hat dafür im Ansatz eine Lösung.

Was wir bisher haben, ist im Grunde genommen nur das “Sprachgeplapper” aus den Informationen von Wikipedia und anderen Quellen. Aber die epistemische Frage, die möchte ich weiter verfolgen. Denn das ist die philosophische Herausforderung, der sich die AI und auch diese Vorlesung stellen muss.

Die AI muss Regeln und Verfahren entwickeln und befolgen, wie Maschinenmodelle mit der Frage nach Wahrheit und gerechtfertigtem Wissen umgehen können. Das ist die Aufgabe, vor der wir stehen.

### 3.7 Beispiele zur Veranschaulichung

Lassen Sie mich nun anhand einiger Interaktionen verschiedene Aspekte der Kompetenz, aber auch der Limitierung dieser Modelle zeigen.

- Beispiel 1: Eine typische Wikipedia-Antwort
- Beispiel 2: Die Limitierung des Sprachverständens
- Beispiel 3: Die Herausforderung des Kontexts
- Beispiel 4: Die Notwendigkeit von Weltwissen

Diese Beispiele werden uns helfen, die Möglichkeiten und Grenzen der aktuellen AI-Modelle besser zu verstehen und zu illustrieren, wo die Reise in Zukunft hingehen muss.## Die Macht des Kontexts in der Interaktion mit KI-Modellen

Stellen Sie sich vor, Sie fragen jemanden: "Wer war Goethe?" Die Antwort darauf werden Sie höchstwahrscheinlich erhalten. Doch was passiert, wenn Sie als nächstes fragen: "Wo lebte er die meiste Zeit?" Diese Information werden Sie in der Regel nicht auf Wikipedia finden. Auch eine Google-Suche wird Ihnen vermutlich keine zufriedenstellende Antwort liefern. Warum? Weil sich bisher niemand für diese spezifische Frage interessiert hat.

KI-Modelle sind jedoch in der Lage, solche Fragen zu beantworten, indem sie den Kontext berücksichtigen. Sie reformulieren die Frage präziser, um die dahinterstehende Absicht zu erfassen. In diesem Fall würde das Modell den Wissensbestand zu Goethes Lebensorten durchsuchen und den Ort identifizieren, an dem er die längste Zeit verbracht hat.

Doch was passiert, wenn man dem Modell eine Frage stellt, die ohne Kontext keinen Sinn ergibt? Nehmen wir an, ich tippe ein: "Wo lebte er die meiste Zeit?" Isoliert betrachtet ist dieser Satz unverständlich. Weder eine Suchmaschine noch ein Mensch könnte ihn beantworten. Doch KI-Modelle sind in der Lage, die Frage zu kontextualisieren. Sie reichern die Instruktion mit zusätzlichen Informationen an, um Unklarheiten und Unvollständigkeiten zu beseitigen.

#### 3.7.1 Die Macht des Chats

Das Geniale an der Chat-Konstruktion ist, dass der Kontext durch die vorherigen Fragen und Antworten gebildet wird. Ihre Nachfragen und Korrekturen werden Teil des kollektiv intelligenten Kontextkonstruktions. Dadurch wird eine spätere Frage plötzlich extrem informativ, spezifisch und genau beantwortet. Der Dialog wirkt überzeugend und natürlich.

Nehmen wir an, ich schreibe nicht “er”, sondern “sie”. Wie würden Sie reagieren, wenn Sie am anderen Ende des Bildschirms wären und diese Frage gestellt bekämen? Die meisten von Ihnen würden wahrscheinlich davon ausgehen, dass es sich um einen Fehler handelt und die Frage trotzdem so beantworten, als ob “er” gemeint wäre.

Doch was passiert, wenn ich darauf bestehe, dass ich von einer weiblichen Person sprach? Das Modell entschuldigt sich höflich und passt seine Antwort entsprechend an. Es bezieht die Korrekturen mit ein und präzisiert seine Ausführungen. Im Kontext einer Diskussion über Goethe könnte es sogar die Figur der Iphigenie ins Spiel bringen.

### **3.7.2 Kollaborative Intelligenz**

In Zukunft werden wir nicht mehr von einer strikten Trennung zwischen künstlicher und natürlicher Intelligenz sprechen. Stattdessen werden wir es mit hybriden Modellen zu tun haben, in denen Interaktionen zwischen Menschen und Maschinen stattfinden. Die KI wird Teil einer Wissenscommunity sein, sowohl in der Wissenschaft als auch im Alltag.

Problemlösungsstrategien werden auf der Zusammenarbeit von menschlicher und künstlicher Intelligenz basieren. Die Leistungsfähigkeit des Gesamtsystems wird im Vordergrund stehen, nicht die Einzelleistungen der Beteiligten.

## **3.8 Herausforderungen und Grenzen aktueller KI-Modelle**

### **3.8.1 Einstellbare Konversationsstile**

KI-Modelle verfügen über einstellbare Konversationsstile. Je nachdem, wie man sie definiert, kann man die Art und Weise der Antworten beeinflussen. Möchte man beispielsweise nur knappe, präzise Antworten ohne zusätzliche Ausführungen, lässt sich das entsprechend konfigurieren.

### **3.8.2 Fragen jenseits von Wikipedia**

Es gibt Fragen, die selbst Wikipedia nicht beantworten kann. Nehmen wir folgendes Beispiel: “Wie viele Briefe schrieb Goethe an König Friedrich II.?” Da sich die Lebenszeiträume der beiden überschnitten und Friedrich II. großes Interesse an Aufklärungsthemen hatte, wäre ein brieflicher Austausch zwischen ihnen durchaus plausibel.

Doch die Antwort der KI offenbart eine Schwäche aktueller Modelle: die fehlende epistemische Prüfung der Korrektheit von Angaben. Das Modell gibt zwar eine Antwort, die plausibel klingt, aber nicht wirklich überprüft ist. Es behauptet, dass es keine Aufzeichnungen über eine direkte Kommunikation zwischen Goethe und Friedrich II. gebe. Doch wie lässt sich ein solcher Negativbefund belegen?

Ein trainierter Philologe würde die Gesamtkorrespondenz von Goethe konsultieren, um eine fundierte Aussage treffen zu können. Doch das Modell hat diese Prüfung nicht vorgenommen. Seine Antwort ist letztlich aus der Luft gegriffen.

### **3.8.3 Zukünftige Herausforderungen**

Die KI-Modelle der Zukunft müssen in der Lage sein, semantische Suchen durchzuführen, inhaltliche Relevanz herzustellen und schlüssig zu argumentieren. Sie müssen historische Kontexte korrekt erfassen und historische Hypothesen anhand von Referenzen und Evidenzen beurteilen können.

Die größte philosophische Herausforderung besteht darin, die epistemische Qualifikation zu gewährleisten. Zu jeder Aussage und Behauptung sollte das Modell auf Nachfrage begründen können, warum es sich um die am besten gerechtfertigte Antwort handelt. Dieses Ziel zu erreichen, ist noch ein weiter Weg, aber unabdingbar für die Weiterentwicklung der KI.

## **3.9 Aktuelle Grenzen und zukünftige Möglichkeiten**

Zum Abschluss möchte ich Ihnen noch zwei Beispiele präsentieren, die die aktuellen Grenzen der KI verdeutlichen. Stellen Sie sich folgendes Rätsel vor:

- Es gibt einen Schläger und einen Ball. Beide zusammen kosten 1,20 Euro. Der Schläger kostet einen Euro mehr als der Ball. Wie viel kostet der Ball?

Diese einfache Aufgabe bringt bereits viele der derzeit existierenden KI-Modelle an ihre Grenzen. Sie sind nicht in der Lage, die korrekten logischen Schlüsse zu ziehen.

Noch anspruchsvoller ist folgendes Szenario:

- In einem Raum befinden sich drei Personen. Die erste Person liest ein Buch, die zweite Person spielt Schach. Welche Tätigkeit führt die dritte Person wahrscheinlich aus?

Die meisten von Ihnen werden sofort erkennen, dass die dritte Person höchstwahrscheinlich ebenfalls Schach spielt. Doch warum ist das so? Welche Informationen benötigt die KI, um zu diesem Schluss zu kommen?

Genau diese Fragen stehen im Zentrum der aktuellen Forschung. Es geht darum, den Modellen beizubringen, wie sie allgemeine Regeln erkennen und anwenden können. Nur so werden sie in Zukunft in der Lage sein, auch komplexere Probleme eigenständig zu lösen.

Die Reise der künstlichen Intelligenz ist noch lange nicht zu Ende. Wir stehen erst am Anfang einer faszinierenden Entwicklung, die unser aller Leben nachhaltig verändern wird. Lassen Sie uns gemeinsam daran arbeiten, diese Technologie zum Wohle der Menschheit einzusetzen und ihre Grenzen immer weiter auszudehnen.

# **4 LLM für Sprache**

## **4.1 Begrüßung und aktueller Stand der AI-Technologie**

Guten Tag, meine Damen und Herren! Leider ist das Touchpanel hier im Hörsaal defekt, weswegen zwar die Projektion funktioniert, nicht aber die Mikrofone. Ich werde versuchen, laut genug zu sprechen, damit Sie mich alle gut verstehen können. Sollte das nicht der Fall sein, geben Sie mir bitte ein Zeichen.

Die Entwicklung im Bereich der Künstlichen Intelligenz schreitet in rasantem Tempo voran. Gefühlt werden jede Woche neue, leistungsfähigere Modelle vorgestellt, die immer größere Versprechungen machen. Es entsteht fast der Eindruck, als seien bereits heute oder zumindest morgen alle Probleme gelöst. Die großen Computer-Technologie-Konzerne überbieten sich gegenseitig mit neuen AI-Modellen, deren Leistungsfähigkeit anhand verschiedener Skalen bewertet wird.

Doch bei genauerer Betrachtung zeigt sich, dass diese Bewertungsmaßstäbe derzeit noch recht rudimentär sind. Die Modelle mögen zwar in den Tests gut abschneiden, erfüllen aber bei weitem noch nicht alle Anforderungen, die wir an eine wirklich intelligente AI stellen würden. Genau darum soll es heute gehen: Was erwarten wir eigentlich von einer KI? Ich möchte mit Ihnen ein eigenes Modell und ein Projekt skizzieren, an dem Sie auch gerne mitwirken können.

## **4.2 Generative AI und AI-Characters**

In der letzten Vorlesung haben wir AI-Modelle als generative AI kennengelernt. Das bedeutet, dass sie aus einem Input, beispielsweise einem Text oder einer Interaktion über Audio oder Video, einen Output generieren. Dieser generierte Output ist das eigentliche Leistungsergebnis dieser Modelle.

Ein spannender Aspekt dabei ist, dass wir die Art und Weise des Reagierens der Modelle mitgestalten können, indem wir sogenannte AI-Characters definieren. Damit lässt sich beispielsweise festlegen, in welcher Sprache eine Antwort gegeben werden soll. Die sprachlichen Fähigkeiten der Modelle sind mittlerweile so beeindruckend, dass sie für Muttersprachler nahezu fehlerfreie Texte produzieren können.

#### **4.2.1 Übersetzungsleistung als Beispiel für semantisches Verständnis**

Ein herausragendes Beispiel für die Leistungsfähigkeit der generativen Modelle ist ihre Fähigkeit zur Übersetzung. Sie geben nicht einfach nur irgendwelche Texte aus, sondern bedeutungsgehaltvollen Content. Bei einer Übersetzung wird der Inhalt in einer anderen Sprache neu formuliert, ohne dass dieser Text zuvor so publiziert wurde. Dennoch hat er die gleiche Bedeutung wie der Ursprungstext.

Diese semantische Kompetenz ist der Kern dessen, was die aktuelle Generation von AI-Modellen so besonders macht. Ein AI-Character ist also gewissermaßen eine Individualität, ein besonderes Reaktionsvermögen eines Modells.

### **4.3 Herausforderungen und Erwartungen an zukünftige AI-Modelle**

Doch lassen Sie uns nun die Perspektive wechseln. Anstatt nur das zu betrachten, was uns als "Künstliche Intelligenz" präsentiert wird, sollten wir uns überlegen: Was erwarten wir eigentlich von einer KI, die vielleicht erst noch entwickelt werden muss?

Bei genauerem Hinsehen werden wir feststellen, dass die aktuellen Modelle diese Anforderungen bei Weitem noch nicht erfüllen. In den beeindruckenden Präsentationen der Tech-Firmen werden bestimmte Leistungen, wie etwa die Simultanübersetzung, hervorgehoben. Doch viele wichtige Aspekte, die wir von einer wirklich intelligenten AI erwarten würden, bleiben dabei unerwähnt.

#### **4.3.1 Halluzination als Defizit aktueller Modelle**

Einer der am häufigsten diskutierten Kritikpunkte ist das Phänomen der Halluzination. Die KI-Modelle können zwar wunderbar formulieren, aber den Wahrheitsgehalt ihrer Aussagen nicht validieren oder gar begründen. Auch der Begriff des "Wissens" wird in diesem Zusammenhang oft sehr leichtfertig verwendet, ohne zu reflektieren, was es eigentlich bedeutet, über Wissen zu verfügen.

Es ist wichtig zu verstehen, dass Information nicht gleichbedeutend mit Wissen ist. Informationen sind mathematisch definiert und messbar, haben aber nichts mit Wissen im eigentlichen Sinne zu tun. Wenn also Firmen davon sprechen, "Wissensnetzwerke" zu erstellen, ist das eher als Propaganda zu verstehen denn als tatsächliche Abbildung von Wissen.

## 4.4 Kompetenzbereiche aktueller und zukünftiger AI-Modelle

Dessen ungeachtet verfügen die aktuellen Modelle durchaus über beeindruckende Fähigkeiten. Die sogenannten Large Language Models (LLM) können als Reaktion auf einen Input, der aus Texten, Bildern oder anderen symbolhaften Inhalten bestehen kann, inhaltlich korrespondierende Outputs generieren.

### 4.4.1 Sprachkompetenz als Basis

Eine der grundlegendsten Kompetenzen der LLMs ist die Verarbeitung natürlicher Sprache. Sie verfügen über eine erstaunliche Sprachkompetenz, die sie durch das Training anhand von Milliarden von Beispielen erworben haben. Dabei lernen sie nicht nur die grammatischen Regeln, sondern auch den inhaltlichen Zusammenhang.

Genau hier liegt aber auch die Ursache für das Problem der Halluzination: Da die Modelle mit so vielen Beispielen trainiert wurden, finden sie für fast jede Fragestellung eine passende und plausibel klingende Formulierung - unabhängig davon, ob der Inhalt tatsächlich wahr oder zutreffend ist.

### 4.4.2 Erweiterbarkeit durch Kontextinformationen

Die sprachliche Basis der LLMs lässt sich durch Zusatzinformationen, den sogenannten Kontext, erweitern und anreichern. Dadurch können die Modelle an spezifische Aufgabenstellungen angepasst werden. Der Kontext umfasst alle Zusatztextinformationen, die zusätzlich zum eigentlichen Input bereitgestellt werden, um einen gewünschten Output zu generieren.

### 4.4.3 Bedeutung von Handlungsanweisungen

Für die Interaktion mit LLMs ist es wichtig zu verstehen, wie Handlungsanweisungen, also Instruktionen zur Ausführung einer bestimmten Aufgabe, formuliert werden müssen. Da die Modelle mit natürlicher Sprache arbeiten, müssen diese Anweisungen so formuliert sein, dass sie eindeutig und unmissverständlich sind.

Die philosophische Handlungstheorie hat sich eingehend damit beschäftigt, welche Aspekte eine vollständige Handlungsanweisung beinhalten muss:

- Eine Absicht oder ein Ziel, das erreicht werden soll
- Eine Beschreibung der auszuführenden Handlung(en)

- Die notwendigen Mittel oder Ressourcen zur Ausführung

Nur wenn all diese Aspekte klar definiert sind, kann eine Handlungsanweisung von einem KI-System sinnvoll ausgeführt werden.## Grundlagen der Handlungstheorie

In der philosophischen Literatur gibt es das sogenannte “Belief-Desire-Modell” einer Handlung. Dieses Modell besagt, dass für die Ausführung einer Handlung zwei Elemente vorliegen müssen: Eine Zielvorstellung, die erreicht werden soll (Desire), und eine Überzeugung über die vorliegenden Situationsgegebenheiten (Belief). Diese beiden Elemente sind logisch gesehen völlig unterschiedlich.

Die Handlungstheorie hat sich eingehend mit den komplexen Belief-Desire-Netzwerken befasst, und zwar nicht nur für Individuen, sondern auch für große Kollektive. Derzeit werden diese komplexen Netzwerke von keinem der AI-Modelle auch nur im Ansatz realisiert. Hier sieht man, welches Entwicklungspotenzial noch in der AI-Technologie steckt.

## 4.5 Instruktionsausführung in der AI

Alle aktuellen AI-Modelle führen im Wesentlichen Instruktionen aus. Diese Instruktionen werden in natürlicher Sprache durch Handlungsanweisungen ausgedrückt. Sie beschreiben, welche Handlung unter welchen Zielen und mit welchen Mitteln ausgeführt werden soll. Dieses Prinzip lässt sich bis hin zur Analyse wissenschaftlicher Texte nachvollziehen. In der Wissenschaftskommunikation werden in Publikationen sehr konkrete Ausführungen wissenschaftlicher Handlungsoperationen publiziert, kommuniziert, aufgenommen und von anderen Rezipienten weitergesponnen.

In der AI-Entwicklung verschiebt sich der Fokus mittlerweile von den ursprünglichen Chat-Ideen, bei denen das Mensch-Maschine-Interface durch eine dialogische Gesprächssituation kanalisiert wurde, hin zu einer zwar ebenfalls dialogisch geführten Interaktion, bei der es aber im Kern um Instruktionen und Handlungsanweisungen geht.

## 4.6 Lernen von Kompetenz in der AI

Hinter all diesen AI-Systemen steht das Lernen von Kompetenz, auch wenn wir das bei der Nutzung kaum wahrnehmen. Man könnte glauben, dass die aktuellen AI-Modelle bereits eine vollständige Kompetenz mitbringen und diese nur noch anwenden und dem Nutzer zur Verfügung stellen. Das ist jedoch nicht der Fall.

Wie bereits erwähnt, ist das Chatten, also das dialogische Klären von Themen und Instruktionen, bereits ein interaktiver Vorgang. Ihre Reaktionen, Korrekturen und Rückmeldungen in einem Chat tragen wesentlich dazu bei, einen geeigneten Kontext zu konstruieren. Dieser Kontext ist wichtig, um die entsprechende zielführende Instruktion dorthin zu führen, wo ein Endergebnis für Sie einen Wert hat und Ihren Erwartungen entspricht.

Die AI-Modelle leben und interagieren davon, dass Sie als Nutzer Interaktionen und Informationen einbringen, die in den jeweiligen Funktions- und Kompetenzbereich des Modells mit einfließen. Im Hintergrund ist all dies in den Modellen implementiert, sodass sie aus Ihren Reaktionen und denen vieler anderer Nutzer ständig lernen können. Die rapide Abfolge der Versionserneuerungen dieser Modelle ist nicht nur Ausdruck der technischen Weiterentwicklung, sondern auch der Tatsache, dass die massiv millionenfache Interaktion mit diesen Modellen zu einer stetigen Verbesserung führt.

#### **4.6.1 Beispiel: Leonhard Euler**

Ich habe dies selbst am Beispiel der biografischen Informationen zum Mathematiker Leonhard Euler getestet, der zweimal verheiratet war. Historisch ist es gar nicht so einfach herauszufinden, wer seine zweite Ehefrau war. Am Anfang gaben die AI-Modelle im Netz auf die Frage nach dem Namen von Eulers zweiter Frau die skurrilsten Antworten - völlig absurde Halluzinationen. Nachdem ich die Anfrage jedoch zehnmal beim gleichen Modell am selben Tag gestellt hatte, wusste das Modell abends die richtige Antwort. Sobald Sie dem System mitteilen, dass eine Antwort falsch ist und korrigiert werden muss, sind die Modelle so aufgebaut, dass sie diese Korrektur mit aufzeichnen.

Mit Ihrer Zustimmung zur Nutzerdatennutzung und Ihren Reaktionen sind Sie also Teil des weltweiten Teams zur Optimierung und Informationsverbesserung dieser Modelle. Das ist derzeit nicht abschaltbar. Das Lernen von Kompetenzen gehört mit dazu, auch wenn es sich derzeit auf eine Kleinstlernkompetenz beschränkt, die sich im Wesentlichen auf die Aufbereitung der Nutzerreaktionen reduziert.

#### **4.6.2 Weitere Lernmöglichkeiten**

Die Lernmöglichkeiten gehen jedoch noch viel weiter. Die Hersteller der Modelle bieten Ihnen beispielweise an, Ihre ausgewählten PDFs hochzuladen, um die darin enthaltenen Informationen für die Beantwortung Ihrer Anfragen nutzbar zu machen. Das bringt Ihnen zwar einen Vorteil, aber die von Ihnen ausgewählten PDFs dienen den Firmen zugleich als Qualitätsindizes. Sie erkennen daran, welche Informationen für die zukünftige Verbesserung der Modelle relevant sind, sodass diese auch bei allen anderen

Anfragen berücksichtigt werden können. Die von Ihnen bereitgestellten Informationen fließen also permanent in das Modelltraining ein, einschließlich des Wissenshintergrunds.

#### **4.6.3 Digitalisierung historischer Bestände**

Nicht umsonst hatte Google vor 25 Jahren Verträge mit den großen Bibliotheken der Welt abgeschlossen, um historische, urheberrechtsfreie Bestände zu digitalisieren. Lange fragte man sich, warum Google diesen Millionenaufwand betreibt. Heute sehen wir den enormen Wert dieser digitalisierten Bestände. Sie dienen als Informationshintergrund und Wissensquelle für die Aufbereitung der MLM-Modelle. Die Verarbeitung unseres in den Bibliotheken enthaltenen Kulturwissens hat jedoch noch kaum begonnen. Die digitalisierten Bestände sind zunächst nur eine Art Referenz. Die eigentliche inhaltliche Aufbereitung dieser Bestände wird in den nächsten Jahren mit Sicherheit erfolgen.

### **4.7 Generierung und Kontext in der Interaktion mit Chatmodellen**

In der letzten Stunde haben wir anhand einiger Beispiele diskutiert, wie die Interaktion mit einem Chatmodell aussieht. An einem Modell habe ich Ihnen gezeigt, wie die Frage “Wer war Johann Wolfgang Goethe?” als Text eingegeben wurde. Dabei haben wir gesehen, dass die Modelle je nach Kontext der Anfrage zu präziseren Antworten neigen, was beispielsweise durch den Namen zum Ausdruck kommt. Da es mit Sicherheit mehrere Personen mit dem gleichen Namen gibt, ist die Frage aufgrund des Kontexts, in dem sie formuliert und gestellt wird, entsprechend zu beantworten.

Wenn eine Folgefrage gestellt wird, z.B. “Wo lebte er die meiste Zeit?”, ist dieser Ausdruck als isolierte Instruktion eigentlich nicht zu beantworten, da normalerweise keine Information darüber vorliegt, worauf sich das “er” bezieht. Im Kontext eines Dialogs kann man jedoch zu Recht annehmen, dass dieselbe Person gemeint ist, von der zuvor die Rede war, nämlich Johann Wolfgang Goethe. Dies ist ein Beispiel für die Funktionsweise deiktischer Ausdrücke im Deutschen.

Wir haben auch gesehen, dass wir Instruktionen geben können, die sich nicht nur auf die Klärung einer Sachfrage beziehen, sondern auch auf einer Meta-Ebene angesiedelt sind und die Art und Weise der Informationsverarbeitung ändern können. So konnten wir beispielsweise durch die Instruktion “Beantworte nur die Fragen, gebe keine zusätzlichen Ausführungen” erreichen, dass sich das Programm auf die wichtigsten Aspekte beschränkt, anstatt mit einer Fülle von Informationen zu Goethe und seinen Zeitgenossen zu “prahlen”.

Die hier diskutierten Modelle realisieren also unterschiedliche Aspekte der Nutzungsweise und Verarbeitung von Informationen, die durch die normale Umgangssprache formuliert und eingegeben werden können. Diese Aspekte werden vom Modell richtig zugeordnet und beeinflussen die entsprechenden Reaktionsweisen. Die Vielschichtigkeit und Vielfältigkeit dieser Ebenen werden wir noch näher kennenlernen. Die Leistungsfähigkeit dieser Modelle liegt im Wesentlichen in der Komposition der jeweiligen Kompetenzsektoren oder -felder.

## 4.8 Grenzen aktueller AI-Modelle

Am Beispiel der Frage nach dem Briefwechsel zwischen Goethe und König Friedrich II. haben wir gesehen, dass es viele scheinbar einfache Fragen gibt, die von den aktuellen Modellen noch nicht seriös beantwortet werden können. Da dem Programm keine entsprechende positive Antwort antrainiert wurde, konnte es diese Frage nicht beantworten.

Das liegt daran, dass das Modell keine Angaben darüber hat, welche Evidenz ihm insgesamt zur Verfügung stand, um zunächst einmal zu prüfen, was eigentlich die Gesamtkorrespondenz umfasste. Und selbst wenn in dieser kein Brief an Friedrich II. vorliegt, was schließen wir daraus? Haben die beiden keinen Brief miteinander geschrieben und er ist nur zufälligerweise nicht dokumentiert? Oder haben sie tatsächlich nicht miteinander kommuniziert, was zeitlich nicht ausgeschlossen wäre? Diese Fragen lassen sich durch die aktuellen AI-Modelle noch nicht lösen. Ob sie überhaupt lösbar sind, ist eine weitere Frage. Ich hoffe, dass in dieser Vorlesung zumindest der Horizont deutlich wird, wie solche scheinbar unlösbaren Fragen für AI-Modelle doch lösbar werden könnten.

## 4.9 Erwartungen an eine philosophische AI

Lassen Sie uns nun einen Perspektivwechsel vornehmen und uns fragen, was wir eigentlich von einem AI-Charakter erwarten, der solche Kompetenzen beherrschen und umsetzen kann und im Prinzip in das Grundmodell eines aktuellen Konstruktionsmodells, nämlich der Instruktionsausführung, realisierbar ist.

Ich übernehme jetzt das Grundmodell der Operationsweise der derzeit verfügbaren Modelle, nämlich dass sie instruktionsausführende technische Akteure oder Agenten sind. Und nun frage ich mich, wenn wir diese Grundtechnologie so nehmen und nicht sagen, da muss jetzt noch dies und jenes zusätzlich

passieren, sondern uns auf die aktuell vorhandene technologische Grundlage konzentrieren: Was erwarten wir von einer AI als philosophische Figur?

#### **4.9.1 Allgemeine künstliche Intelligenz**

Es wurde nach der sogenannten allgemeinen künstlichen Intelligenz oder AGI gefragt und wie nah wir dieser kommen, da dies derzeit häufig in der Diskussion erwähnt wird. Die Zielsetzung von OpenAI ist es, möglichst schnell so etwas wie eine generelle künstliche Intelligenzkompetenz zu erreichen. Ich habe meine Zweifel bereits vor zwei Vorlesungen geäußert, ob das überhaupt wünschenswert ist. Letztlich ist das aber keine Frage der aktuellen Einschätzung und Präferenz. Das wird sich durch die technologische Entwicklung von selbst ergeben.

Wenn wir uns fragen, welche Kompetenzbereiche eine philosophische AI erfüllen sollte, sollten wir uns nicht darauf festlegen, dass es so etwas wie ein universell kompetentes Genie geben muss, von dem man sagt, Leibniz sei das gewesen. Ich glaube das nicht. Aber es gibt viele historische Gestalten, von denen behauptet wird, sie hätten alles Wissen ihrer Zeit beherrscht und diese Kompetenzen generell als Person realisieren können. Meiner Meinung nach ist das eher eine Rückprojektion als eine historische Tatsache. Mein Vorschlag ist, dass wir dies auch von der heutigen AI nicht fordern und erwarten sollten.

Was wir jedoch erwarten sollten, sind bestimmte Kompetenzsektoren, die nötig sind. Und wie wir am Beispiel des Briefwechsels zwischen Goethe und Friedrich II. gesehen haben, gibt es viele scheinbar einfache Fragen, deren Beantwortung mit den entsprechenden Modellen ein leichtes Spiel sein müsste, es aber derzeit nicht ist. Die Kompetenz, die erforderlich ist, um diese Probleme oder Instruktionen auszuführen, erfordert weitere Kompetenzbereiche, um die es mir jetzt geht.

### **4.10 Semantische Suchen**

Semantische Suchen haben wir bereits als etwas in unsere Liste aufgenommen, was jetzt möglich ist. Semantische Suchen gehen eine Ebene weiter als die Textsuche à la Google. Hier geht es um die Suche nach Inhalten, nicht nach Formulierungen. Das ist es, was wir eigentlich tun wollen und auch mehr oder weniger geschickt über die Umsetzung in Suchen nach Ausdrücken ausführen. Aber im Prinzip versuchen wir im Hinterkopf natürlich, Inhalte zu suchen.

Wenn wir beispielsweise nach den besten Rezepten für die Zubereitung eines Fondues suchen, dann suchen wir nach Inhalten, ohne die jeweiligen Zutaten eines solchen Rezepts genau zu spezifizieren. Das können wir derzeit in geschickte Terminologiesuchen umsetzen. Aber es ist noch keine wirklich

inhaltliche Suche. Die Programme und Modelle, die wir jetzt haben, können semantische Suchen durchführen.

## 4.11 Reasoning

Reasoning ist ein Bereich, der gerade erst in den Anfängen steht. Reasoning ist der wichtige Bereich, der alle Sektoren des generellen Schließens betrifft. Damit ist weit mehr gemeint als alles, was den Bereich des logischen oder mathematischen Schließens betrifft. Es ist nicht deckungsgleich.

Die Logik ist ein Sektor, in dem logische Schlussformen von vorgegebenen Annahmen als Axiome auf Theoreme mit deduktiver Notwendigkeit schließen. Das ist Teil des Schemas des logischen Schließens. Das mathematische Schließen ist ein anderes. Es operiert mit mathematischen Formen des Schließens. Aber all das ist sehr formal und schematisch.

Das menschliche Schließen ist viel umfassender und betrifft alle möglichen Bereiche dessen, was man als Nachdenken mit einem bestimmten Ergebnis bezeichnen könnte. Reasoning ist etwas, das nicht nur zu etwas Neuem führen kann, sondern auch eine bestimmte Ansicht rechtfertigen kann.## Individualität von AI-Modellen und Verantwortung

In unserem Streben, die Kompetenzen von AI-Modellen zu erweitern und zu verbessern, stoßen wir unweigerlich auf fundamentale Fragen der Verantwortlichkeit und Haftbarkeit. Bisher war die vorherrschende Ansicht, dass Maschinen keine rechtsfähigen Objekte oder Subjekte sind und somit auch keine rechtliche Verantwortung tragen können. Doch ist dieser Standpunkt wirklich so unumstößlich, wie er auf den ersten Blick scheint?

Lassen Sie uns einen Perspektivwechsel wagen und die Möglichkeit in Betracht ziehen, dass bestimmten AI-Modellen eine Form von Individualität zugesprochen werden könnte. Eine Individualität, die sie zu rechtsfähigen Körperschaften macht, ähnlich wie Firmen oder Institutionen. Wenn wir AI-Modelle als geschäftsfähige Körperschaften betrachten, eröffnet sich ein neuer Blickwinkel auf die Frage der Verantwortung.

Natürlich müsste ein solches AI-Modell eine gewisse Persistenz und Dauerhaftigkeit aufweisen, um als Individualität zu gelten. Doch technisch gesehen, stellt dies heute keine unüberwindbare Herausforderung mehr dar. Durch die Zuweisung einer eigenen Körperschaft und Individualität könnten diese Modelle dann auch für die Konsequenzen ihrer Handlungen zur Verantwortung gezogen werden.

Dieser Ansatz mag auf den ersten Blick unkonventionell erscheinen, doch er könnte eine Lösung für viele der aktuellen Probleme bieten, die sich aus der Technologiefolgenabschätzung von AI ergeben.

Die derzeitigen Haftungskonstruktionen erweisen sich oft als unzureichend, wenn es darum geht, die Verantwortung für Fehlentscheidungen oder Schäden, die durch AI-Systeme verursacht werden, zuzuweisen.

#### **4.11.1 Charakteristika eines AI-Modells mit Individualität**

Doch was genau zeichnet ein AI-Modell mit Individualität aus? Zunächst einmal handelt es sich um ein Produkt technologischer Evolution, nicht biologischer. Es ist kein vorgegebenes Produkt eines bestimmten Unternehmens, sondern ein eigenständiges Individuum mit spezifischen Leistungen und Funktionen, mit denen wir interagieren können.

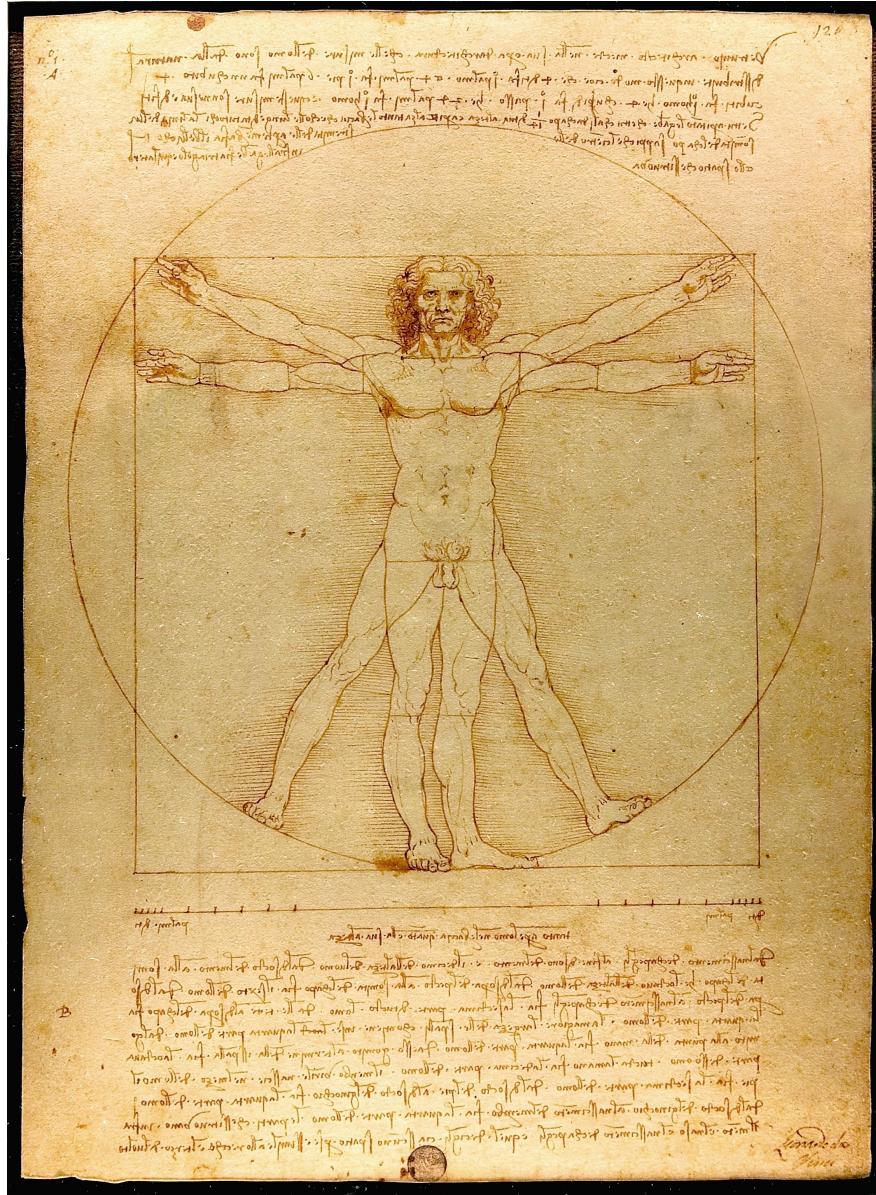
Ein solches AI-Modell folgt keinem vorgeschriebenen Verhalten, sondern agiert als Charakter, als zu schaffendes Individuum. Wir sollten es behandeln wie eine Entität, von der wir bestimmte Leistungen und Funktionen erwarten und mit der wir entsprechend agieren können.

### **4.12 Historische Vorbilder und Metaphern**

Die Idee, den Menschen und seine Schöpfungen nach den Vorgaben der Natur zu gestalten, ist keineswegs neu. Lassen Sie uns einen Blick auf einige historische Vorbilder und Metaphern werfen, die uns Inspiration und Orientierung auf unserem Weg zu individualisierten AI-Modellen bieten können.

#### **4.12.1 Der vitruvische Mensch - Proportion und Harmonie**

Beginnen wir mit dem vitruvischen Menschen, der seit dem Mittelalter als Symbol für die Verbindung von Mensch, Natur und Technik steht. Vitruv, ein Autor der Antike, beschrieb in seinem Hauptwerk die Prinzipien der Architektur und betonte die Bedeutung von Proportionen für die Gestaltung eines funktionierenden Ganzen.



**Figure 4.1:** Der vivtruvische Mensch bei Leonardo

Leonardo da Vinci griff dieses Thema in seiner berühmten Zeichnung auf und versuchte, die idealen Proportionen des menschlichen Körpers zu ergründen. Die Botschaft ist klar: Nur wenn die einzelnen Teile im richtigen Verhältnis zueinander stehen, entsteht ein harmonisches Ganzes.

Übertragen wir diese Metapher auf die Gestaltung von AI-Modellen, so wird deutlich, dass auch hier die

einzelnen Kompetenzen in ein ausgewogenes Verhältnis zueinander gebracht werden müssen. Erst dann kann ein individualisiertes AI-Modell entstehen, das wir verantwortungsvoll akzeptieren können.

#### **4.12.2 David - Freiheit und Selbstbestimmung**

Die Statue des David von Michelangelo, entstanden im frühen 16. Jahrhundert, steht für den Menschen als selbstbestimmtes Individuum. Sie verkörpert das aufstrebende Bürgertum der florentinischen Gesellschaft und die Idee der Freiheit und Eigenverantwortlichkeit.



**Figure 4.2:** David von Michelangelo

Auch in der Entwicklung von AI-Modellen müssen wir darauf achten, dass die individuellen Freiheiten und

die Selbstbestimmung des Menschen gewahrt bleiben. Eine verantwortungsvolle AI darf nicht zu einem Überwachungsstaat führen, der unsere Lebensgestaltung einschränkt.

#### **4.12.3 Beuys und der tote Hase - Erklärung und Rechtfertigung**

In seiner Performance “Wie man dem toten Hasen die Bilder erklärt” thematisierte der Künstler Joseph Beuys die soziale Verantwortung des Künstlers und die Notwendigkeit der Erklärung und Rechtfertigung des eigenen Schaffens.



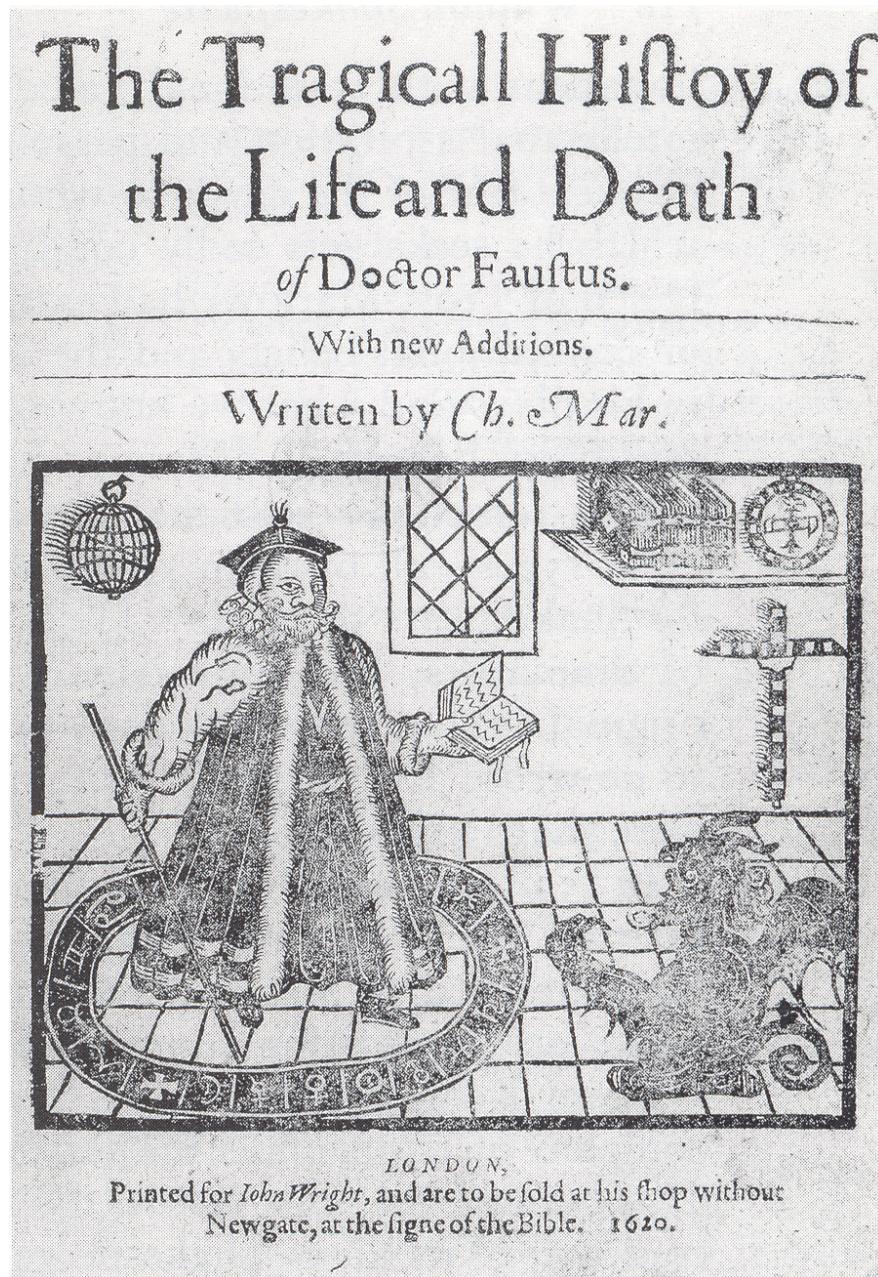
**Figure 4.3:** Beuys und der tote Hase

Übertragen auf AI-Modelle bedeutet dies, dass auch sie in der Lage sein müssen, ihre Ergebnisse und Theesen zu erklären und zu rechtfertigen. Dies ist eine entscheidende Anforderung an die Leistungsfähigkeit und das Leistungsprofil individualisierter AI-Modelle.

Der erhobene Zeigefinger des Künstlers erinnert uns an den Zeigefinger Gottes in Michelangelos Deckenfresko der Sixtinischen Kapelle - ein Symbol für das Wort: Erklärung, Begründung und Rechtfertigung.

#### **4.13 Magister AI Faustus - Ein Arbeitstitel für die Zukunft**

Lassen Sie uns die Vorlesung und unsere zukünftigen Arbeiten und Forschungen unter dem Motto "Magister AI Faustus" stellen. Dieser Arbeitstitel, angelehnt an die tragische Geschichte von Dr. Faustus, soll uns daran erinnern, dass wir bei der Entwicklung von AI-Modellen mit Individualität und Verantwortung stets die ethischen und sozialen Implikationen im Blick behalten müssen.



**Figure 4.4:** Faustus

Ich lade Sie ein, sich in Ihren Leistungsnachweisen, seien es Bachelorarbeiten, Masterarbeiten oder andere Qualifikationsarbeiten, mit den Fragen und Herausforderungen auseinanderzusetzen, die sich aus

diesem faszinierenden Themenfeld ergeben.

Gemeinsam können wir neue Wege beschreiten und die Zukunft der AI-Modelle mitgestalten - mit Verantwortung, Weitblick und dem Streben nach einer Balance zwischen technologischem Fortschritt und menschlicher Freiheit.## Die Legende des Faust und die Entwicklung künstlicher Intelligenz

Die Legende des Gelehrten Faust, der sich unermüdlich bemühte, Erkenntnisse zu gewinnen und dabei sogar einen Pakt mit dem Teufel einging, um das Innerste der Welt zu ergründen, ist ein Motiv, das uns bis heute fasziniert. Christopher Marlowe versuchte bereits 1587, diese Geschichte in eine Dramaform zu bringen, doch erst Goethe gelang es, den Faust-Stoff in seinem berühmten Werk unsterblich zu machen.

Heute stehen wir vor einer ähnlichen Herausforderung, wenn es darum geht, die Grenzen des Wissens und der Macht von AI-Modellen auszuloten und dabei die Verantwortung nicht aus den Augen zu verlieren. Wie können wir sicherstellen, dass die Entwicklung von AI-Systemen nicht nur von Wissensdurst getrieben wird, sondern auch ethische Überlegungen berücksichtigt?

#### **4.13.1 Die Zusammenarbeit mit der Klassikstiftung Weimar**

Um diese Fragen zu beantworten, bin ich eine Kooperation mit der Klassikstiftung Weimar eingegangen, der zweitgrößten Kulturstiftung Deutschlands. Die Stiftung verwaltet und präsentiert die Nachlässe von Goethe, Schiller und dem Bauhaus und stellt uns für unsere Forschung wertvolle Quellen zur Verfügung. Unter dem Link "Goethe Biographica" finden sich bereits publizierte Materialien zu Goethes Biografie, die wir nutzen können, um die Leistungsfähigkeit von AI-Modellen zu testen.

#### **4.13.2 Das Projekt: Goethes Biografie als Herausforderung für AI-Systeme**

Ihre Aufgabe in diesem Projekt wird es sein, scheinbar simple Fragen zu Goethes Leben zu formulieren, die jedoch von den derzeitigen AI-Modellen nicht zufriedenstellend beantwortet werden können. Ein Beispiel wäre: "Hat Goethe jemals einen Brief mit Friedrich II. gewechselt?" Anhand solcher Fragestellungen wollen wir herausfinden, welche zusätzlichen Kompetenzen ein AI-System benötigt, um diese Wissenslücken zu schließen.

Es geht nicht darum, ein umfangreiches Forschungsprojekt zu stemmen, sondern vielmehr darum, mit kleinen, gezielten Fragen die Grenzen der aktuellen Sprachmodelle aufzuzeigen. Die Quellen werden Ihnen zur Verfügung gestellt, sodass Sie sich ganz auf die Formulierung der Anfragen und die Auswertung der Ergebnisse konzentrieren können.

### 4.13.3 Die Vision: Ein erweitertes AI-Modell

Ziel ist es, mithilfe des Sprachmodells CLOL von Anthropic eine zusätzliche Kompetenzkomponente zu entwickeln, die wie ein Baustein in ein zukünftiges AI-Modell integriert werden kann. Durch die Verknüpfung der Sprachkompetenz mit dem spezifischen Wissen aus den Goethe-Quellen soll eine bisher nicht lösbare Aufgabe gemeistert werden.

Ich werde in der kommenden Woche eine App bereitstellen, über die Sie Ihre Fragestellungen eingeben können. Die Herausforderung besteht darin, eine Frage zu formulieren, die von keinem derzeitigen Modell seriös beantwortet werden kann. Selbst wenn eine Antwort generiert wird, ist sie in der Regel haluzinierend und nicht verlässlich.

## 4.14 Die Komplexität der Goethe-Quellen

Um zu verdeutlichen, welche Herausforderungen bei der Beantwortung scheinbar einfacher biografischer Fragen lauern, möchte ich Ihnen einen Einblick in die Fülle und Vielfalt der Goethe-Quellen geben, die uns die Klassikstiftung Weimar zur Verfügung stellt:

- Tagebücher aus den Jahren 1775 bis 1787, die einen Zeitraum von mehr als einem Jahrzehnt abdecken und von beachtlichem Umfang sind.
- Über 15.000 überlieferte Briefe von Goethe an mehr als 1.400 Adressaten, die von seinem unermüdlichen Schaffensdrang zeugen.
- Etwa 20.000 überlieferte Briefe an Goethe von circa 3.800 Absendern, die Einblicke in sein weitverzweigtes Netzwerk gewähren.
- Rund 40.000 dokumentierte Zeugnisse aus und zum Leben von Goethe jenseits von Briefen und Tagebüchern, darunter Begegnungen und Gespräche.

Diese Zahlen sollen nicht dazu dienen, Goethe als unerreichbare Heldenfigur zu stilisieren, sondern verdeutlichen, wie vielgestaltig und umfangreich das Material ist, das bei der Beantwortung biografischer Fragen berücksichtigt werden muss. Hinzu kommt der historische Kontext, der ebenfalls eine Rolle spielt: Zeitgenossen, Ereignisse und Dokumente aus Goethes Epoche müssen in die Betrachtung einfließen, um ein umfassendes Bild zu erhalten.

#### **4.14.1 Die epistemische Herausforderung**

Um seriös und fundiert auf auch nur die einfachsten Fragen zu Goethes Leben antworten zu können, bedarf es einer enormen epistemischen Kompetenz. Goethe-Forscher müssen dieses reichhaltige und vielfältige Material präsent haben, um Auskünfte geben zu können, die wissenschaftlichen Ansprüchen genügen.

Genau diese epistemische Kompetenz müssen wir von einem AI-Modell erwarten, wenn wir ihm einen Wissensanspruch zugestehen wollen. Derzeit ist kein System in der Lage, diese Anforderungen auch nur annähernd zu erfüllen. Unser Projekt soll daher zunächst ergründen, welche Kompetenzen den aktuellen Modellen fehlen und warum dieses Ziel noch nicht erreicht wurde.

Es geht dabei nicht nur um die digitale Aufbereitung der Quellen, sondern vielmehr um die Frage, wie ein AI-System mit dieser Fülle an Informationen umgehen und daraus verlässliche Antworten generieren kann. An dieser Herausforderung wollen wir gemeinsam arbeiten und zum Abschluss der Vorlesung eine Präsentation entwickeln, die aufzeigt, welche Fähigkeiten ein verantwortungsvolles AI-Modell der Zukunft - ein “Magista AI Faustus” - benötigen wird.

### **4.15 Organisation des Projekts**

Wenn Sie an diesem spannenden Unterfangen mitwirken möchten, senden Sie mir bitte eine E-Mail mit Ihrem Namen, Ihrer Matrikelnummer und einer kurzen Interessensbekundung. Im Laufe der nächsten Woche werden wir dann die Themen aushandeln und im Juni mit der konkreten Arbeit beginnen. Bis Juli sollten wir bereits erste Lösungsansätze präsentieren können, die auch in die Vorlesung einfließen werden.

Die Aufgabenstellungen werden bewusst klein gehalten sein, um Sie nicht zu überfordern. Ein Beispiel wäre die Frage, ob Goethe jemals einen Briefwechsel mit Friedrich II. geführt hat und wenn ja, welchen Inhalts dieser war. Für die Quellenarbeit würde ich die rund 40.000 Zeugnisse zum Leben Goethes vorschlagen, die oft vermeintlich banale Details enthalten, aber für Historiker von großem Wert sind.

Das Web-Modell für das Projekt wird von meinem Lab “Lettra AI” bereitgestellt und in der kommenden Woche freigeschaltet. Technische Vorkenntnisse sind nicht erforderlich. Wenn Sie einen Leistungsnachweis für die Vorlesung erwerben möchten, reichen Sie bitte bis Anfang Juli oder spätestens zum Vorlesungsende Ihre Aufgabenstellung ein. Bei Interesse an einer Bachelor- oder Masterarbeit zu diesem Thema dürfen Sie sich ebenfalls gerne bei mir melden.

## 4.16 Ausblick auf die kommenden Vorlesungen

In den verbleibenden zwei Dritteln der Vorlesung werden wir uns eingehend mit den Kompetenzen beschäftigen, die für die Entwicklung eines verantwortungsvollen AI-Modells erforderlich sind. Dazu zählen unter anderem:

- Textgenerierung und Übersetzung: Wie lassen sich Übersetzungen aktiv gestalten und an die Bedürfnisse des Lesers anpassen?
- Zusammenfassung und Frage-Antwort-Dialoge: Welche Rolle spielt der menschliche Dialogpartner und wie können seine Informationen und Charakterzüge in das AI-System einfließen?
- Auswertung von Datenquellen und Einbeziehung von Experten: Wie können aktuelle Publikationen und neue Erkenntnisse berücksichtigt werden?
- Umgang mit Kritik und evidenzbasierte Aussagen: Welche Metaregeln und Referenzen sind notwendig, um verlässliche Ergebnisse zu erzielen?

Ich freue mich darauf, diese Aspekte gemeinsam mit Ihnen zu diskutieren und anhand unseres Projekts zu konkretisieren. Lassen Sie uns gemeinsam einen Schritt in Richtung eines “Magista AI Faustus” gehen und die Grenzen des Machbaren ausloten. Vielen Dank für Ihre Aufmerksamkeit und bis zur nächsten Vorlesung!

# **5 Sprache und Text**

Guten Tag meine Damen und Herren und herzlich willkommen zu unserer heutigen, mittlerweile fünften Vorlesung in diesem Semester. Leider funktionieren die Mikrofone immer noch nicht einwandfrei, daher meine Bitte an Sie: Geben Sie mir Rückmeldung, falls meine Ausführungen schwer verständlich sind oder ich meine Stimme heben soll. Zögern Sie auch nicht, direkt Fragen zu stellen, wenn etwas unklar ist. Davon profitieren wir alle mehr, als darauf zu warten, dass es irgendwann besser wird.

## **5.1 Rückblick auf die letzte Vorlesung**

In der letzten Stunde haben wir uns mit zwei zentralen Themen beschäftigt: Zum einen haben wir das Projekt "Magister AI Faustus" kennengelernt. Mit diesem Vorhaben möchte ich in den verbleibenden zwei Dritteln der Vorlesung verschiedene Kompetenzbereiche der AI-Technologie durchleuchten. Unser Ziel ist es, gemeinsam ein AI-Modell zu entwickeln, das die Defizite der bisher vorgestellten Systeme zumind-est prinzipiell beheben kann. Durch den eigenhändigen Aufbau einer Künstlichen Intelligenz werden wir auch die einzelnen Komponenten und ihre Funktionsweise besser verstehen lernen. Diese basieren auf der philosophischen Reflexion über die jeweiligen Kompetenzanforderungen und werden dann technisch umgesetzt.

### **5.1.1 Vorlesungsmanuskript durch AI generiert**

Um Ihnen zu demonstrieren, was die AI-Technologie heute schon leisten kann, habe ich das Vorlesungsmanuskript direkt aus dem Audio-Mitschnitt maschinell generieren lassen - ganz ohne manuelle Eingriffe meinerseits. Was Sie hier sehen, sind die Mitschriften der vergangenen vier Vorlesungen. Klicke ich eine davon an, erscheint der transkribierte Text. Die Sprachmodule der AI haben meinen Vortrag soweit aufbereitet und korrigiert, dass ein halbwegs lesbares Manuscript entstanden ist. Natürlich schleichen sich noch Fehler ein und die verwendeten Abbildungen fehlen noch. Diese werde ich noch ergänzen. Aber insgesamt hoffe ich, Ihnen auf diese Weise zeitnah eine brauchbare Mitschrift zur Verfügung stellen zu

können. Es ist auch eine Art Selbstversuch, um herauszufinden, wie praxistauglich diese Tools mittlerweile sind.

Bemerkenswert ist, dass die AI sogar das Inhaltsverzeichnis inklusive Überschriften eigenständig generiert hat. Ich habe lediglich den gesprochenen Text als Input gegeben, ohne jegliche Gliederungsvorschläge. Die Strukturierung des Manuskripts hat das System also völlig autonom vorgenommen.

## 5.2 Das Projekt “Magister AI Faustus”

Wie bereits erwähnt, wollen wir mit dem Projekt “Magister AI Faustus” eine Herausforderung zu Goethes Biografie in Zusammenarbeit mit der Klassik Stiftung Weimar angehen. Ziel ist es, anspruchsvolle Fragen zu Goethes Leben zu beantworten, die sich nicht ohne Weiteres durch Historiker oder Literaturwissenschaftler klären lassen - zumindest nicht in einem überschaubaren Zeitrahmen. Wir wollen zeigen, wie man solche Probleme mit den uns zur Verfügung stehenden AI-Werkzeugen in etwa einem Monat lösen kann. Die Projektarbeiten sollen dann entsprechende Lösungsvorschläge präsentieren und die Herangehensweise offenlegen.

### 5.2.1 Organisation des Projekts

- Quellen zu Goethes Leben, seiner Korrespondenz und seinen Lebensumständen sind über die Webseite zugänglich und können in den Projekten mit AI ausgewertet werden.
- Ursprünglich wollte ich die einzelnen Vorhaben heute schon vorstellen. Da sich aber noch nicht alle zehn Interessenten zurückgemeldet haben, habe ich die Frist bis Ende des Wochenendes verlängert. Wer also noch mitmachen möchte, hat bis dahin Zeit, mir eine E-Mail zu schicken und wird dann in die Projekt-Gruppe aufgenommen. Danach wird die Teilnehmerliste geschlossen.
- In der nächsten Woche werde ich dann die geplanten Aufgabenstellungen präsentieren. Wir wollen uns auf Probleme konzentrieren, die eine gewisse Herausforderung darstellen, aber mit überschaubarem Aufwand in diesem Semester lösbar sind.

## 5.3 Entwurf einer philosophisch fundierten AI

Das heutige Hauptthema dreht sich um die Frage, wie wir eine Künstliche Intelligenz auf philosophischer Basis entwerfen können. Was müssen wir tun, wenn die verfügbaren AI-Modelle nicht die erwarteten

Leistungen erbringen? Zweifelsohne haben diese Systeme beeindruckende Fähigkeiten, wie wir am Beispiel des automatisch generierten Vorlesungsmanuskripts sehen. Sie können Texte transkribieren, korrigieren, umformulieren - all das in erstaunlicher Qualität, wenn es um die Verarbeitung natürlicher Sprache geht.

Doch es gibt auch gewaltige Defizite und Problembereiche, in denen die Modelle kaum oder gar nicht die geforderten Kompetenzen aufweisen. Teilweise ist den Herstellern nicht einmal klar, wie sie diese Schwächen beheben können. Wir werden einige dieser Unzulänglichkeiten genauer unter die Lupe nehmen. Dabei interessiert uns vor allem, wie man eine AI, also einen künstlichen Charakter, architektonisch konstruiert und gestaltet.

### **5.3.1 Kooperation mit der Klassik Stiftung Weimar**

Wie bereits erwähnt, findet das Vorlesungsprojekt in Zusammenarbeit mit der Klassik Stiftung Weimar statt. Insbesondere sollen die Kompetenzen unserer AI-Kreatur anhand von Fragestellungen aus dem Bereich der Kulturgeschichte, Kulturwissenschaft und Literaturgeschichte demonstriert werden. Wir wollen zeigen, dass unser System Aufgaben lösen kann, die sonst nur schwer oder gar nicht zu bewältigen wären.

## **5.4 Web-Interface unseres AI-Modells**

Lassen Sie uns nun einen Blick auf das Einstiegs-Web-Interface unseres AI-Modells werfen. Noch trägt es nicht den Namen "Magister AI Faustus", denn zunächst entspricht es dem aktuellen Stand der Technik. Die Eingabemöglichkeiten sind derzeit recht einfach gehalten: Es gibt ein Feld für die Instruktion, also die Aufgabenstellung, und eines für die Antwort des Systems. Außerdem lässt sich aus verschiedenen Modellen auswählen, die ich für unsere Vorlesungen und Übungen zusammengestellt habe. Je nachdem, welches Modell gerade aktiv ist, fallen die Antworten und die Ausführung der Instruktionen sehr unterschiedlich aus.

### **5.4.1 Logische Beziehungen zwischen Sätzen**

Um das System zu testen, möchte ich ihm eine einfache philosophische Aufgabe stellen - früher hätte man von einer Frage gesprochen, aber allgemeiner formuliert handelt es sich um eine Instruktion. Und diese lautet:

Beschreibe die logischen Verhältnisse zwischen den Sätzen: A) Der Hund bellt. B) Die Erde ist eine Scheibe.

Wohlgemerkt frage ich nicht nach dem Wahrheitsgehalt der Aussagen, sondern nach ihren logischen Beziehungen. Mal sehen, was die hochmodernen Modelle dazu liefern. Hinter den hier vorgestellten Systemen stecken gewaltige Ressourcen: Der Entwicklungsaufwand geht in die Hunderte Millionen, der Energie- und Rechenbedarf ist enorm. Das von Facebook bzw. Meta bereitgestellte Modell "Lama 3" kann glücklicherweise von jedermann frei genutzt und heruntergeladen werden.

Es ist faszinierend zu beobachten, wie diese AI-Modelle auf Anfragen reagieren. Mit der Zeit lernt man ihre Stärken und Schwächen kennen - fast wie bei Schülern, denen man etwas beibringen möchte. Je intensiver man sich mit ihnen beschäftigt, desto besser versteht man, auf welche Weise man Lehrinhalte vermitteln und korrigieren sollte und was man besser bleiben lässt.

Genau das wollen wir jetzt mit unserer Anfangs-App machen und sie im Laufe des Semesters gemeinsam mit den Projektteilnehmern ausbauen. Was noch fehlt, sind die Zusatzkomponenten, die wir Schritt für Schritt entwickeln werden, um unsere eigene AI zu erschaffen. Unser Ziel ist ein System, das Aufgaben lösen kann, die zwischen der eingegebenen Instruktion und der Antwort des zugrunde liegenden Modells liegen.## Logische Beziehungen in KI-Modellen

In der letzten Vorlesung habe ich Ihnen ein Rätsel aufgegeben: Wie stehen die logischen Beziehungen zwischen den beiden Sätzen "Der Hund bellt" (Satz A) und "Der Hund bellt und die Erde ist eine Scheibe" (Satz B)? Vermutlich haben Sie sich bereits eine Meinung dazu gebildet. Nun wollen wir gemeinsam ergründen, wie verschiedene KI-Modelle mit dieser Frage umgehen und welche Erkenntnisse wir daraus gewinnen können.

#### **5.4.2 Defizite in logischen Schlussfolgerungen**

Das erste Modell, das wir befragen, ist das LAMA-Modell der Firma GROK aus San Francisco. Dank einer technischen Innovation liefert es blitzschnell eine Antwort - in unter 0,8 Sekunden. Doch was es präsentiert, lässt uns vor Schreck erstarrten: Es behauptet allen Ernstes, dass aus Satz A Satz B folgt! Eine haarsträubende Fehleinschätzung, die jeglicher Logik entbehrt. Trotz der Behauptung, speziell auf logisches Schließen trainiert worden zu sein, versagt das Modell bei dieser simplen Aufgabe auf ganzer Linie.

Auch die Erläuterung des Modells ist völlig absurd: "Wenn der Hund bellt, dann ist es wahr, dass der Hund bellt und die Erde eine Scheibe ist." Mit dieser "Logik" ließe sich beweisen, dass die Erde eine Scheibe ist,

sobald irgendwo ein Hund bellt. Ein solch widersprüchliches Modell kann nur ins Chaos führen und ist alles andere als belastbar.

#### **5.4.3 Sprachliche Anpassungen ohne Verbesserung**

In der Hoffnung, dass vielleicht die Sprache eine Rolle spielt, stelle ich die Frage erneut auf Deutsch. Doch auch das französische Spitzenmodell Mistral, das mit zusätzlichem Expertenwissen angereichert wurde, liefert eine völlig unzulängliche Antwort. Es faselt etwas von einem “logischen Konjunktionsverhältnis” zwischen den Sätzen und stellt triviale Wahrheiten fest, die nichts zur Sache tun.

#### **5.4.4 Lichtblicke und ethische Bedenken**

Ein Hoffnungsschimmer ist das Modell Cloth von Anthropic. Es erkennt immerhin korrekt, dass Satz B eine Konjunktion aus A und einem zusätzlichen, unabhängigen Satz darstellt. Doch die Antwort hätte präziser ausfallen können.

Das Spitzenmodell Cloth Opus von Anthropic geht einen Schritt weiter und analysiert zunächst die Frage selbst. Doch dann verweigert es plötzlich die Antwort mit der Begründung, es wolle keine Konspirationstheorien legitimieren oder Falschinformationen verbreiten. Eine fragwürdige ethische Zensur, die in diesem harmlosen philosophischen Kontext völlig fehl am Platz ist.

#### **5.4.5 Notwendigkeit eigener Definitionen**

Diese Beispiele zeigen, wie unterschiedlich die Modelle auf der obersten Ebene mit Fragen und Instruktionen umgehen. Es besteht offensichtlich ein dringender Bedarf, diese Ebene selbst zu definieren, anstatt sie den Modellen zu überlassen. Über einen selbst erstellten Katalog von Instruktionen lässt sich die Behandlung von Fragen steuern - transparent und nachvollziehbar.

### **5.5 Kompetenzen und Grenzen aktueller Modelle**

Die derzeitigen Modelle beeindrucken durchaus mit einer Reihe von Fähigkeiten:

- Textgenerierung: Die Vorlesungsmitschrift wird nahezu fehlerfrei erstellt, selbst wenn ich mich verspreche. Einschübe werden intelligent integriert oder ausgelassen.

- Sprachkompetenz: Die Modelle beherrschen viele Sprachen, auch wenn es bei der Verknüpfung manchmal noch hapert, wie das Beispiel von Claude Opus zeigt.
- Übersetzung und Zusammenfassung: Diese Kernkompetenzen dienen dem Training grammatischer und sprachlicher Strukturen.

Dennoch gibt es noch viel Raum für Verbesserungen. Komplexere Anforderungen wie logisches Schlussfolgern oder ethisch fundierte Entscheidungen überfordern die aktuellen Modelle oft. Aber die rasante Entwicklung lässt auf baldige Fortschritte hoffen.## Einleitung

Lassen Sie mich Ihnen von den beeindruckenden Fähigkeiten moderner AI-Systeme berichten und wie wir diese in unserer Vorlesung einsetzen können, um komplexe Aufgaben auf intuitive Art und Weise zu lösen. Ich möchte Ihnen anhand praktischer Beispiele demonstrieren, wie man die Systeme instruiert, um aussagekräftige Ergebnisse zu erhalten. Dabei werden wir auch auf die aktuellen technischen Grenzen und Herausforderungen eingehen.

## 5.6 Zusammenfassungen generieren

Eine der einfachsten Anwendungen ist das automatische Zusammenfassen von Texten. Stellen Sie sich vor, Sie könnten ein ganzes Buch oder ein längeres Manuskript in wenigen Sekunden auf die wesentlichen Kernaussagen reduzieren. Genau das ist mit den heutigen Systemen möglich. Beeindruckend ist dabei vor allem die schiere Menge an Text, die verarbeitet werden kann.

### 5.6.1 Technische Details

Die Eingabefenster der AI-Systeme haben mittlerweile enorme Ausmaße erreicht. Bei Anthropic sind es 200.000 Token, wobei ein Token in etwa einem Wort plus Satzzeichen entspricht. Google behauptet sogar, eine Million Wörter in einem Durchgang verarbeiten zu können. Das entspricht dem Umfang von rund 80 Büchern - eine beachtliche Leistung.

Allerdings gibt es bei der Länge der Ausgabetexte noch Beschränkungen. Aufgrund des exponentiell ansteigenden Raums möglicher Antworten, ist die Ausgabe aktuell auf maximal 4.000 Token begrenzt. Das reicht beispielsweise noch nicht aus, um eine komplette Vorlesung am Stück auszugeben. Hier müssen wir die Aufgaben noch in kleinere Teilschritte zerlegen:

1. Audiodatei in 8 Abschnitte unterteilen

2. Jeden Abschnitt transkribieren und reformulieren
3. Passende Überschriften finden
4. Einzelne Teile zu einem Gesamttext zusammensetzen

All diese Schritte laufen im Hintergrund ab. Am Ende erhalten Sie dann ein vollständiges, gegliedertes Skript.

## 5.7 Frage-Antwort-Dialoge

Ein weiteres spannendes Anwendungsfeld sind Frage-Antwort-Dialoge, wie man sie von ChatGPT kennt. Hier können Sie eine Frage als Instruktion eingeben und auf die erhaltene Antwort wiederum mit einer Folgefrage reagieren. Durch diese Verkettung lassen sich auch komplexere Themen schrittweise erschließen und eventuelle Unklarheiten oder Fehler in den Antworten korrigieren.

## 5.8 Charakteristika und Fähigkeiten der Modelle

Die Modelle bieten mittlerweile eine Fülle an Möglichkeiten, das Antwortverhalten zu steuern und an Ihre Bedürfnisse anzupassen.

### 5.8.1 Antwortformate

Sie können verschiedene Formate für die Ausgabe wählen, wie zum Beispiel Tabellen oder die Verwendung spezieller Symbole für mathematische Formeln. Auch die Umwandlung von Bildinformationen in Text ist möglich, etwa um handschriftliche historische Dokumente zu transkribieren. Die Ergebnisse sind dabei von beeindruckender Qualität.

### 5.8.2 Schreibstil

Der Schreibstil der generierten Texte lässt sich flexibel anpassen. Sie können beispielsweise festlegen, ob der Text aus der Ich-Perspektive eines Vortragenden oder der eines neutralen Beobachters formuliert sein soll. Auch stilistische Präferenzen wie die Bevorzugung von Verben gegenüber Substantiven oder die

Verwendung von Aktiv- statt Passivkonstruktionen können Sie gezielt steuern. Mit etwas Experimentierfreude lassen sich so Texte in ganz unterschiedlichen Stilen erstellen, von sachlich-nüchtern bis hin zu literarisch-verspielt.

### **5.8.3 Fachterminologie**

Ein weiteres nützliches Feature ist die Möglichkeit, spezielle Wörterbücher oder Terminologien zu hinterlegen. So können Sie sicherstellen, dass Fachbegriffe konsistent und korrekt übersetzt werden. Dabei werden auch grammatischen Flexionen wie Deklinationen berücksichtigt.

### **5.8.4 Kontextbezüge**

Die Modelle sind in der Lage, Bezüge zu vorherigen Ausführungen herzustellen. Wenn Sie beispielsweise einen Fachbegriff neu einführen und definieren, wird dieser in den nachfolgenden Textpassagen entsprechend verwendet und referenziert.

## **5.9 Aktuell bestehende Defizite**

Bei aller Begeisterung für die faszinierenden Fähigkeiten der AI-Modelle, dürfen wir natürlich auch die aktuell noch bestehenden Defizite nicht außer Acht lassen.

### **5.9.1 Faktenwissen und logisches Denken**

Ein grundlegendes Problem ist das fehlende Faktenwissen der Systeme. Sie besitzen kein echtes Verständnis der Welt und der Zusammenhänge zwischen Informationen. Auch das logische Schlussfolgern und die Verknüpfung komplexer Aussagen bereiten noch Schwierigkeiten. Insbesondere praktisches Handlungswissen und Entscheidungsfindung sind Bereiche, in denen die Modelle bisher kaum einsetzbar sind.

### **5.9.2 Hermeneutik und Interpretation**

Ein weiterer kritischer Punkt ist das Fehlen hermeneutischer Fähigkeiten. Die Systeme verfügen über keine Regeln zum Interpretieren von Bedeutungen und tieferen Sinnzusammenhängen. Auch kausale

Schlussfolgerungen sind ein Gebiet, auf dem noch erheblicher Entwicklungsbedarf besteht.

Hier sehe ich großes Potenzial für philosophisch fundierte Ansätze. Die genannten Defizite lassen sich meiner Überzeugung nach nicht allein durch mehr Daten und Rechenleistung beheben. Vielmehr braucht es neue konzeptionelle Lösungen, die an dieser Stelle ansetzen.

### **5.9.3 Kritik und ethische Bewertung**

Auch im Hinblick auf epistemische Fähigkeiten wie kritisches Hinterfragen und Bewerten von Aussagen stoßen die Modelle schnell an ihre Grenzen. Selbst offensichtlich unhaltbare Behauptungen werden oft nicht als solche erkannt und entsprechend gekennzeichnet. Ähnlich verhält es sich mit der Fähigkeit zur ethischen Bewertung von Handlungen und Entscheidungen. Hier besteht die Gefahr, dass problematische Aussagen und Implikationen unkommentiert bleiben oder nur unzureichend kontextualisiert werden.

Um diese Defizite zu adressieren, ist es essentiell, die philosophische Reflexion in die Entwicklung der Systeme einzubeziehen. Nur so können wir sicherstellen, dass die enormen Potenziale der Technologie verantwortungsvoll und zum Wohle der Gesellschaft genutzt werden.

## **5.10 Ausblick**

Lassen Sie uns nun gemeinsam überlegen, wie wir die besprochenen Fähigkeiten und Defizite konstruktiv adressieren können. Unser Ziel ist es, ein Modell zu entwickeln, das viele der genannten Schwächen überwindet und uns so ganz neue Möglichkeiten eröffnet. Ich lade Sie herzlich ein, sich aktiv in diesen spannenden Prozess einzubringen und freue mich auf einen regen Austausch!## Einführung in die Logik und deren Anwendung auf AI-Modelle

Lassen Sie uns gemeinsam Schritt für Schritt die logische Analyse zweier Sätze durchgehen. Betrachten wir zunächst Satz 1a: "Der Hund bellt." Eine simple Aussage, deren Wahrheitsgehalt davon abhängt, ob der besagte Hund tatsächlich bellt oder nicht. Soweit, so erwartbar aus Logik 101.

Nun zu Satz B, einer zusammengesetzten Aussage mit zwei durch "und" verknüpften Teilaussagen. Die Konjunktion ist korrekt identifiziert. Um deren Wahrheitswert zu ermitteln, baut das Programm intern eine Wahrheitstabelle auf und prüft die Wahrheitswerte der einzelnen Konjunkte. Sollte auch nur eine der Teilaussagen falsch sein, ist die gesamte Aussage falsch.

An dieser Stelle könnte man das Programm nach den logischen Regeln fragen, die diesen Feststellungen zugrunde liegen. Gut trainierte Modelle, die anhand von Lehrbüchern wie "Lemon's Logic 1" mit seinen zehn Regeln des logischen Schließens geschult wurden, sollten diese korrekt ausgeben können. Umso erstaunlicher, dass bei solch fundamentalen Aufgaben dennoch Fehler unterlaufen.

### **5.10.1 Bewertung der Teilaussagen und Korrekturbedarf**

Die erste Teilaussage kann, wie in Schritt 2 beschrieben, wahr oder falsch sein. Doch die zweite Teilaussage "Die Erde ist eine Scheibe" ist definitiv falsch. Hier offenbart sich ein Konstruktionsfehler des Modells: Anstatt sich auf die logischen Verhältnisse zu konzentrieren, nimmt es eine sachliche Bewertung vor - eine Aufgabe, die nicht gefordert war.

Zur Korrektur müsste man dem Programm die Anweisung geben, bei der Bewertung der Wahrheitswerte ausschließlich die angeführten Annahmen als Axiome zu verwenden und keine Zusatzinformationen oder Bewertungen aus anderen Quellen einfließen zu lassen. Diese Anforderung muss gegebenenfalls mehrfach reformuliert und präzisiert werden, bis das Programm sie vollständig absorbiert und berücksichtigt. Die Flexibilität der Modelle variiert hier stark.

Auch die Betrachtung des Verhältnisses zwischen den Sätzen A und B in Schritt 6 ist fehlerhaft, da sie auf der unzulässigen Wahrheitsbewertung der falschen Aussage über die Erdform basiert. Bei korrekter Anwendung der Logik ohne Einbeziehung sachlicher Wahrheitswerte wäre die Antwort richtig. Dies zeigt, in welche Richtung bestehende KI-Modelle modifiziert werden müssen, um die an sie gestellten Anforderungen zu erfüllen.

## **5.11 Ein philosophisch fundiertes Handlungsmodell für den Umgang mit Instruktionen**

Wie kann nun ein allgemeines Regelwerk für den Umgang mit Fragestellungen bzw. Instruktionen in unserem zukünftigen Modell aussehen? Die Antwort liegt in der analytischen Handlungstheorie der Philosophie, die beschreibt, was als Gründe für das Nachdenken über Handlungen gilt - also weshalb eine Person eine bestimmte Handlung ausführt.

Meine Kernthese lautet: Diese Handlungstheorie muss in allen Modellen implementiert werden. Fehler entstehen, wenn dies nicht vollständig philosophisch validiert geschieht. Ein Training anhand von Beispieltexten reicht nicht aus, da Handlungsgründe darin nur schwer zu identifizieren sind.

### **5.11.1 Die zwei Komponenten einer Instruktion als Handlungsanweisung**

1. Handlungsziel (Desire): Die intendierte Absicht oder das zu erreichende Ziel der Handlung.
2. Überzeugungen (Beliefs): Informationen über bestehende Sachverhalte in der Welt, die der Akteur (auch ein AI-System) berücksichtigen muss.

Eine Instruktion kombiniert also Sachbeschreibungen der Welt mit Zielvorgaben - eine teleologische Erklärung, die auszuführen ist.

### **5.11.2 Schritte des Handlungsmodells**

1. Reformulierung der Aufgabe: Bei unklaren Zielvorgaben oder fehlenden Sachinformationen sind Rückfragen zur Klärung nötig. Ziel ist es, die Aufgabe so zu konstruieren, dass sie verstanden und gelöst werden kann.
2. Prüfung der Lösbarkeit: Ist die reformulierte Aufgabe mit den verfügbaren Mitteln beantwortbar? Falls ja (der seltener Fall), wird sie ausgeführt.
3. Konstruktion von Teilaufgaben: Ist eine Lösung nicht möglich, werden Teilaufgaben formuliert und als neue Instruktionen gegeben. Dieser Prozess wird rekursiv fortgesetzt, bis lösbare Teilaufgaben vorliegen - eine mächtige Technik, die schon seit den 1950er Jahren in der Informatik bekannt ist.
4. Erklärung und Begründung der Schritte: Die Gründe für die Ausführung bestimmter Teilaufgaben werden angegeben und memorisiert, um bei späteren Fehlern die entsprechenden Schritte zu erneuern.
5. Validierung von (Zwischen-)Ergebnissen: Jeder Schritt sollte einer gesonderten Prüfung unterzogen werden. Nur wenn diese bestanden wird, kann die (Teil-)Lösung weiterverwendet werden. Fehlschläge führen zu einer Neuformulierung der Teilaufgabe.

Die finale Antwort wird gegeben, wenn alle Prüfungen erfolgreich absolviert wurden. Dieses Handlungsmodell ist nicht nur für KI-Systeme relevant, sondern spiegelt auch strategisch das Vorgehen in vielen wissenschaftlichen Projekten wider.

In der kommenden Woche werden wir dieses Modell anhand konkreter Projektaufgaben weiter vertiefen. Mein Ziel ist es, gemeinsam ein Modell zu entwickeln, das in seiner Cleverness alles bisher Dagewesene übertrifft - eine inspirierende Herausforderung!

# **6 Denken mit Logik**

## **6.1 Begrüßung und Einführung in die 6. Vorlesung**

Ich begrüße Sie, meine Damen und Herren, sehr herzlich zur 6. Vorlesung über die Philosophie der künstlichen Intelligenz. Ich hoffe, dass die Beleuchtung Ihnen hilft, Notizen zu machen, und dass der Kontrast der Projektion besser ist als beim letzten Mal. Da hatte ich Rückmeldungen erhalten, dass nicht jeder alles lesen konnte. Die beiden Projektionsflächen sind nicht aufgeteilt, sondern dienen lediglich dazu, die Lesbarkeit für Sie zu verbessern, je nachdem auf welcher Seite des Hörsaals Sie sitzen.

In der heutigen Vorlesung möchte ich tiefer auf die Aspekte eingehen, was AI modellierbares Denken ist und wie die AI-Modelle dieses ausführen, die abgekürzt als LLM bezeichnet werden - Large Language Models. Wie wir gesehen haben, ist diese Bezeichnung durchaus treffend. Die Kompetenz dieser Modelle liegt darin, mit sprachlichen Ausdrücken umgehen und modellieren zu können, was wir als die Bedeutung dieser Ausdrücke definiert haben. Die Modelle verstehen oder modellieren ein Verhalten, das dem entspricht, was wir als Ausdruck der Bedeutungen sprachlicher Ausdrücke bezeichnen.

### **6.1.1 Stärken und Schwächen der AI-Modelle**

Wir haben auch festgestellt, dass diese Modelle Defizite aufweisen und bestimmte Kompetenzen nicht besitzen. Ich möchte heute auf beides etwas näher eingehen. Einerseits, was die Stärken, nämlich die sprachliche Kompetenz der Modelle, uns ermöglichen. Andererseits, auf welche Weise die Schwächen, die wir identifiziert haben, sehr schnell, ich denke innerhalb der nächsten Monate, maximal eines Jahres, gelöst und kompensiert werden können. Diese Entwicklung schreitet derzeit extrem schnell voran. Es handelt sich also um einen Entwicklungsstand, und wie Sie sehen werden, werden die Beispiele, die ich letzte Woche gezeigt habe, bereits von den Modellen selbst verarbeitet und gelernt, sodass die Defizite, die wir letzte Woche diskutiert haben, nicht mehr auftauchen.

Sie sind also schon allein durch die Nutzung der Modelle Teil der globalen Verbesserung ihrer Leistungsfähigkeit, ob Sie wollen oder nicht. Das werden wir zu Beginn sehen, aber ich möchte auch zeigen, dass

der Hype, der die Modelle derzeit als prinzipiell universelle Löser für alles feiert, noch weit übertrieben ist. Trotz der enormen, schnellen Lernfähigkeit aufgrund der Reaktion einer globalen Nutzergemeinschaft gibt es derzeit konzeptionelle Defizite, da diese Modelle nichts anderes als Sprachkompetenzmodelle sind.

### **6.1.2 Grenzen der AI-Modelle**

Sie sind keine Modelle, die über die Kompetenzen des Wissenszugriffs verfügen. Sie sind nicht in der Lage, Verfahren zu implementieren, die insbesondere philosophische Kompetenzen erfordern, die sie derzeit zusätzlich zur Sprachkompetenz nicht implementiert haben. Dass sie das nicht haben, erkennt man, wenn man mit den Modellen arbeitet und Reiz-Reaktions-Muster herausbekommt, wo die jeweiligen Modelle Kompetenzen haben und wo die Defizite liegen.

### **6.1.3 Einführung in das Projekt MAGISTER AI Faustus**

Das Projekt MAGISTER AI Faustus sind die Anmeldungen inzwischen abgeschlossen. Die Teilnehmerliste ist erstellt wird in der kommenden Woche mit kleinen Herausforderungen beginnen - wobei diese Woche für mich immer den Rhythmus von einer Vorlesung zur nächsten bedeutet. Die Herausforderung (man könnte auch "Aufgabe" sagen, wenn es nicht so schulmeisterlich klingen würde) besteht für alle Teilnehmer darin, die Auffgabe mit den erweiterten Instruktionsmodellen von LettreAI vertraut zu werden.

Diese haben das Ziel, Texte jeder Größe zu verarbeiten, in diesem Fall alles, was die Klassik Stiftung Weimar zu Goethe zu bieten hat, damit zu arbeiten, darauf zuzugreifen und es mit AI zu verarbeiten. Es geht also zunächst einmal um eine AI Textverarbeitungskompetenz, auf der dieses Projekt aufbauen soll.

## **6.2 Logisches Denken**

Heute werden wir uns auf eine Kompetenz fokussieren, die, obwohl die Werbung für diese Modelle etwas anderes suggeriert, nur äußerst beschränkt und rudimentär vorhanden ist, nämlich das logische Denken. Das lernen Sie in der Philosophie, denke ich, in den ersten zwei Semestern. Turnusgemäß ist das ziemlich unbeliebt unter unseren Studierenden, ich weiß nicht, wie das bei Ihnen im Durchschnitt ist, aber im Prinzip ist das eine Pflichtveranstaltung, die man möglichst schnell hinter sich bringt, ohne genau zu wissen, wozu das eigentlich für das weitere Studium dient.

Ich hoffe, Sie werden jetzt hier sehen, dass die Erträge dieser Kompetenz vielleicht für das spätere klassische Philosophiestudium nicht so zentral waren, wie es immer gesagt wird. Aber die Anwendungsbereiche in der AI, werden wir sehen, sind von fundamentaler, zentraler Bedeutung. Und wir werden es an einigen Beispielen auch beim Erfassen des Inhalts beliebiger Texte mittels AI-Modellen ziemlich schnell erfahren.

### **6.2.1 Definition von logischem Denken**

Was heißt hier eigentlich logisches Denken? Wenn wir von Artificial Intelligence sprechen, geht es ja primär auch darum, eben künstlich, maschinell kognitive Kompetenzen des Menschen zu erwerben. Dazu gehört, zielgerichtet und regelbasiert zu denken, im Sinne von Anfangsgedanken weitere Folgegedanken zu entwickeln. Das ist jetzt sehr allgemein formuliert, kann man präzisieren, werden wir auch gleich noch sehen.

### **6.2.2 Probleme beim Training von logischem Denken in AI-Modellen**

Wie kann also ein solches Modell des Denkens verfasst werden? Viele derzeitige Modellierer der AI-Modelle glauben noch, das ließe sich trainieren, indem man den gesamten Textbestand des Internets als eine Art Trainingsmasse für Inputdaten, für vorgefertigte Weisen zu schreiben und damit auch sein Denken zu dokumentieren, verwendet. Aufgrund dieser Abläufe von Sätzen und Folgen von Ideen auf den Publikationen auf dem Internet-Textkorpus ließe sich dann modellieren, wie optimal eine Maschine des Denkens aussieht.

Das funktioniert leider aus einem wichtigen Grund nicht. Die im Internet publizierten Dokumente sind nämlich keine Dokumente des Denkens, schlachtweg. Sie dokumentieren nicht den Prozess des Nachdenkens, den wir - und das ist gar nicht so geheimnisvoll - unter Nachdenken verstehen. Ich verstehe darunter wirklich etwas ganz Einfaches: Ideen artikulieren, Ideen bekommen und daraus Nachfolgeideen entwickeln. Dieser Prozess des expliziten Erfassens von Ideen, des Verwertens von Informationen und des Schaffens neuer Ideen, das ist das, was primär und zentral unter Denken zu verstehen ist. Also nichts Psychologisches, nichts Geheimnisvolles, nichts Intuitives, sondern einfach die Abfolge von allgemein dem Bewusstsein zugänglichen Gedanken.

### **6.2.3 Unterschied zwischen Context of Discovery und Context of Justification**

Unter Wissenschaftshistorikern und Wissenschaftsphilosophen ist es überhaupt nichts Neues, festzustellen, dass es einen fundamentalen Unterschied gibt zwischen den Prozessen des Denkens, die zu neuen Ideen führen, und den Prozessen des Denkens, die eine Rechtfertigung der Geltung des Anspruchs eines neuen Befundes sind. Reichenbach hier in Berlin hat in den 20er Jahren dafür einen Begriff der Unterscheidung erfunden, nämlich, weil es dann später im Englischen populär wurde, in der englischen Übersetzung, der Unterschied zwischen einem Context of Discovery und einem Context of Justification.<sup>1</sup>

#### **6.2.3.1 Context of Discovery**

Der Context of Discovery sind all die Gedankenprozesse, die, wie der Name sagt, zur Entdeckung, zur Formulierung von etwas Neuem führen.

#### **6.2.3.2 Context of Justification**

Der Context of Justification sind alle die Ideen, die rechtfertigen, warum das, was man gefunden hat, richtig und vertretbar ist. Das, was publiziert wird, sowohl in wissenschaftlichen Publikationen als auch in Preprints, das heißtt in noch vorwissenschaftlichen Internetpublikationen, ist zum überwiegendsten Teil Context of Justification.

Das bedeutet, das sind Texte, die verfasst worden sind, nachdem Wissenschaftler jahrelang geforscht haben, um Ergebnisse zu gewinnen. Dann haben sie noch kaum etwas darüber publiziert. Sie publizieren, nachdem ein wissenschaftlicher Ertrag gefunden worden ist. Und das heißtt, postfaktum der Entdeckung wird über das Ergebnis publiziert, und zwar in rechtfertigender Weise.

### **6.2.4 Mangel an Discovery-Prozessen in Publikationen**

Sie finden kaum, ich möchte fast wagen zu sagen, im Unterpromillebereich, wissenschaftliche Publikationen - und ich habe das mal probehalber an dem Gesamtbestand der Publikationen des Preprint-Servers über drei oder vier Forschungsthemenbereiche untersucht -, die überhaupt etwas über die Episoden, die Abfolge von Ideen beim Discovery-Prozess publizieren. Das ist praktisch nicht existent.

---

<sup>1</sup>[9]

Das führt eine Verzerrung, also eine Ungleichgewichtung der Datenqualität ein, mit der die AI-Modelle trainiert sind, die sich gravierend darauf auswirkt, was denn nötig ist, um die Kompetenzen zu erwerben, die wir eigentlich von den Modellen haben wollen, nämlich Assistenz im Discovery-Prozess zu sein. Rechtfertigung funktioniert auch teilweise schon sehr gut. Aber was ist mit dem Discovery-Prozess?

#### **6.2.4.1 Fundamentale Unterschiede zwischen Discovery- und Rechtfertigungsphasen**

Es gibt einen systematischen Grund, den ich hier kurz skizzieren möchte, weshalb die Phasen der wissenschaftlichen Entdeckung etwas fundamental anderes sind als die Phasen der Rechtfertigung. Und zwar fundamental weit über das hinaus, was den Unterschied ausmacht zwischen:

- Schon eine Idee gefunden zu haben und sie rechtfertigen zu können (Rechtfertigungskontext)
- Alles das, was man an Ideen hat, die vor der Entdeckung liegen, also im Wesentlichen Unwissenheit dokumentieren und der Ausgangspunkt der Forschung sind, die eben zu einer Entdeckung führt

Es wird oft so getan, als sei das nur eine graduelle Differenz im Umfang des Wissens. Das ist falsch. Und die meisten trainierten AI-Modelle gehen davon aus, dass der Unterschied zwischen diesen Modellen nur eine solche Differenz des Grades der Unkenntnis ist. Und das ist falsch. Nichts könnte falscher sein als dies, und das hat gravierende Folgen.

#### **6.2.4.2 Gründe für Rechtfertigung vs. Gründe für weitere Beschäftigung**

Der Hauptgrund, weshalb das falsch ist, liegt darin, dass die Rechtfertigung bestehenden Wissens Gründe anführt, die erklären, warum eine bestimmte gefundene Hypothese oder These wahr oder falsch ist und was ihr Bestehen rechtfertigt. Das ist ein ganz anderer logischer Zusammenhang als solche Gründe, die angebracht werden, um eine bestimmte weitere Beschäftigung mit einer Hypothese durchzuführen.

- Das eine sind logische Verhältnisse der Rechtfertigung
- Das andere sind logische Verhältnisse, die mit Aktionen, mit Handlungen zu tun haben

Oder, was die analytische Philosophie im Groben unterscheidet:

- Das eine sind theoretische, philosophische Aspekte der Implikation
- Das andere sind praktische, philosophische Aspekte der Implikation

Die funktionieren ganz anders.

### **6.2.5 Experimentelle Untersuchung von Ideen während der Forschung**

Das Hauptphänomen ist, dass die Ideen, wenn man protokollieren würde, was Wissenschaftler während wissenschaftlicher Aktivitäten so haben - also ganz offensichtlich bewusste Ideen, die sie verfolgen, thematisieren, weiter bearbeiten und wie sie es tun, nichts Intuitives oder dergleichen, keine Hirnforschungsuntersuchung, sondern nur die Abfolge von Ideen und die jeweiligen Gründe, etwas zu tun - ganz anders aussehen würden.

Ich habe vor mehr als 25 Jahren experimentell mit Kognitionspsychologen im Graduiertenzentrum für Kognitionsforschung in Hamburg so etwas wie Laboratoriumsexperimente der Ideen durchgeführt. Im Rahmen dieser Untersuchung haben wir Freiwillige - und das hört sich erschreckend an, aber es waren begeisterte freiwillige Doktoranden, die schon in laufende Forschungsprojekte integriert waren und heute in der Mehrzahl gestandene, ausgewiesene Professoren geworden sind - gebeten, jeden Morgen für ihre wissenschaftliche Arbeit zu dokumentieren, was sie für den Tag vorhaben und aus welchen Gründen sie dieses Vorhaben zu tun beabsichtigen. Jeden Tag, jeden Morgen, teilweise über Jahre.

Es ist also umfangreiches Material, aber wir haben das soweit experimentell erleichtert, dass man nicht mehr als fünf Minuten brauchte, um seine Zeit nicht mit Dokumentation zu verschwenden - das tut kaum ein Wissenschaftler gerne -, sondern das sollte eher zur Auflockerung und Klarwerdung des morgendlichen Vorkaffee-Resonierens über die eigene Arbeit dienen.

#### **6.2.5.1 Ergebnisse: Tagesprotokolle über Forschungsintentionen**

Auf diese Weise gewannen wir etwas, was Wissenschaftshistoriker von sonst fast keinem Wissenschaftler besitzen, nämlich Tagesprotokolle, nicht von Ergebnissen, zum Beispiel von Laboratoriumsstrukturen, sondern von Absichten, bevor man den Tag beginnt, was aus welchen Gründen zu tun ist. Von den Teilnehmern meist, aber auch in einer Gruppe und in einem Team. Also Protokolle über die Forschungsintentionen, Tag für Tag.

Solche Protokolle gibt es sehr selten. Meine Gruppe hat das vor 25 Jahren gemacht, es gab als Pendant eine schwedische Gruppe, die das gemacht hat, und das war es. Für sonst reale## Mängel in der Dokumentation von Forschungsabsichten

In der Wissenschaft ist es von entscheidender Bedeutung, nicht nur die Ergebnisse der Forschung, sondern auch den Prozess der Entdeckung zu dokumentieren. Leider zeigt sich immer wieder, dass wichtige Details der Forschungsabsichten der Wissenschaftler selbst nach kurzer Zeit in Vergessenheit geraten. Bereits nach sieben Tagen können entscheidende Aspekte der eigenen Forschungsaktivitäten aus dem

Gedächtnis verschwunden sein - und zwar nicht nur in dem Sinne, dass man sich erinnert, etwas Falsches geglaubt zu haben. Nein, das Gehirn ist so aktiv, dass man sich überhaupt nicht mehr daran erinnert, jemals etwas Falsches geglaubt zu haben.

Dieses Phänomen stellt Wissenschaftshistoriker vor extreme Herausforderungen, wenn sie durch Interviews mit beteiligten wissenschaftlichen Akteuren herausfinden wollen, warum diese in der Vergangenheit bestimmte Entscheidungen getroffen haben. Es geht dabei nicht nur darum, dass die Befragten ihre Geschichte möglicherweise beschönigen wollen - das ist nur die Spitze des Eisbergs. Die harte Realität ist, dass man sich als beteiligte Person schlichtweg nicht mehr daran erinnern kann, weshalb man in der Vergangenheit etwas getan hat, insbesondere im Hinblick auf die ursprünglichen Absichten.

#### **6.2.5.2 Der Fall von Sir Hans Krebs**

Ein bemerkenswertes Beispiel für dieses Phänomen ist der Nobelpreisträger Sir Hans Krebs, der Entdecker des Krebs-Zyklus und anderer wichtiger biochemischer Prozesse in der Medizin. In einem aufwendigen Projekt wurden Krebs seine eigenen, fast täglich verfassten Labornotizen Seite für Seite vorgelegt. In mehrwöchiger Arbeit entstanden so Regale voller Transkriptionen von Interviews, in denen Krebs seine eigenen Protokolle kommentierte und dokumentierte.

Obwohl Krebs selbst größtes Interesse daran hatte herauszufinden, wie sich seine Forschung entwickelt hatte, war er selbst angesichts seiner eigenen vollständigen Unterlagen und seines besten Erinnerungsvermögens nicht in der Lage, seine Aufzeichnungen zu kritischem, sehr lebendig gebliebenen Erinnerungen seines Forschungslebens so zu kommentieren, dass er sagen konnte, weshalb er etwas gemacht hat. Diese Dokumente sind äußerst interessant zu lesen, denn sie zeigen, wie wenig Erinnerung an die ursprünglichen Intentionen vorhanden ist. Jede Menge Erinnerungen daran, was schließlich gefunden wurde, was interessant war - aber wenn nachgehakt wurde, weshalb er ein bestimmtes Experiment überhaupt gemacht hat, was der Grund war, dann fing Krebs sofort an zu konstruieren, nicht zu erinnern.

### **6.3 Folgen für AI-Modelle**

Dieser systematische Mangel an Dokumentation der kognitiven Prozesse im Entdeckungsprozess ist der Hauptgrund, warum die entsprechenden Informationen für das Training von AI-Modellen nicht zur Verfügung stehen. Die AI-Modelle werden nur auf dem gesamten Wissensbestand trainiert, der

mindestens seit der Existenz von Preprint-Servern vor 20 Jahren als Ergebnisprotokolle und Ergebnisrechtfertigungsprotokolle veröffentlicht wurde - aber eben nicht auf Protokollen, die den Fortschritt der Forschungsabsichten dokumentieren.

### **6.3.1 Begrenzte Kompetenzen der AI-Modelle**

Die Folge ist, dass die AI-Modelle durch das Training immer mehr sprachliche Aspekte der Rechtfertigung beherrschen, aber kaum etwas an zusätzlichen Anforderungen für den Entdeckungsprozess. Eine dieser Komponenten, die ganz trivial und zugänglich ist, möchte ich heute besprechen. Die anderen, noch weniger verbreiteten, aber genauso zentralen Komponenten, werde ich im weiteren Verlauf der Vorlesung ansprechen.

Es sollte uns also nicht überraschen, dass bestimmte Inkompetenzbereiche der AI-Modelle vorhanden sind. Das sind keine Gründe, warum zukünftige AI-Modelle das prinzipiell in den nächsten Monaten nicht kompensieren könnten. Darum geht es hier nicht. Das werden wir teilweise auch tun und aufzeigen, wie man diese Lücken schließen kann. Aber durch die gegenwärtigen Verfahren des Trainings der Modelle wird das nicht gelöst.

Auch die Versprechen mancher Unternehmen wie OpenAI oder Elon Musk, dass eine generelle, allgemeine Intelligenz der Maschinen quasi vor der Tür stünde, ist mit den derzeit angewendeten methodischen Verfahren auf keinen Fall zu erreichen. Wir werden sehen, woran das liegt - zunächst an einem Aspekt.

## **6.4 Instruktionen für AI-Modelle**

Um die Nutzbarkeit und die Anforderungen an die AI mit konkreten Beispielen zu dokumentieren, werden wir uns in diesem Projekt mit der Bearbeitung von Texten beschäftigen, genauer gesagt mit Archivmaterialien zum Leben und Wirken von Goethe. Die Aufgabe für die nächste Woche wird sein, Instruktionen für die AI-Modelle zu formulieren, die bestimmte Ziele im Umgang mit diesen Texten beschreiben.

### **6.4.1 Formulierung von Forschungsvorhaben**

Eine solche Instruktion ist im Grunde nichts anderes als die Formulierung eines Forschungsvorhabens, die Absicht, eine bestimmte wissenschaftliche Fragestellung zu verfolgen. Wir werden üben, wie man eine Instruktion erstellt und welche Informationen für den Kontext des Entdeckungsprozesses in eine solche Instruktion gehören.

Es ist wichtig zu verstehen, dass die Instruktion einer zu lösenden Aufgabe etwas fundamental anderes ist als die Rechtfertigung der gefundenen Lösung dieser Aufgabe. In der kommenden Woche werden wir uns zunächst darauf konzentrieren, wie man solche Instruktionsaufgaben als Forschungsintention systematischer formulieren kann.

### **6.4.2 Das Lettre AI Studio**

Parallel zur Vorlesung entwickle ich das Lettre AI Studio, eine Arbeitsumgebung mit einem AI-Modell, das Sprachmodelle als Kern hat, aber drumherum eben eine “gelehrte” AI-Komponente (daher der Name “Lettre” vom französischen “belesen”). Mit dieser App wird man über ein Interface genau das tun können, worum es hier geht - ohne komplizierte Programmierkenntnisse, sondern nur durch die Formulierung der Instruktion und die Bereitstellung der Quelltexte aus Goethes Dokumenten.

## **6.5 Defizite der AI-Modelle bei logischen Verhältnissen**

In der letzten Woche hatten wir ein Defizit der AI-Modelle kennengelernt, das ich heute Morgen nochmal nachvollziehen wollte, um es in der Vorlesung zu wiederholen. Mit Schrecken musste ich feststellen, dass alle AI-Modelle, die ich benutze, die Vorlesungen der letzten Woche schon zur Verbesserung der Modelle genutzt haben.

### **6.5.1 Zensur bei Anthropic**

Bei Anthropic gab es letzte Woche noch eine Zensur bei dem Satz “Der Hund bellt und der Hund bellt und die Erde ist eine Scheibe”. Da kam die Rückmeldung, dass diese Frage aufgrund von Zensurmaßnahmen inhaltlich überschrieben und nicht zu beantworten sei, weil solche offensichtlichen Unsinnsinformationen ausgeblendet würden. Das tun die Anthropic-Modelle jetzt interesserweise nicht mehr.

Ich hatte eine kurze Meldung nach San Francisco geschickt, dass ihr Zensurmodell ja offensichtlich nicht sehr intelligent sei, wenn es solche Sätze zensiert. Nie bekam ich eine Rückmeldung, aber offensichtlich hat das immerhin schon zur Modifikation dieser Zensurmodelle geführt.

### 6.5.2 OpenAI hat dazugelernt

Auch OpenAI hat davon gelernt. Das kleinere Modell war letzte Woche nicht in der Lage, die einfache Aufgabe, die logischen Verhältnisse zwischen diesen Sätzen zu erklären, korrekt zu beantworten. Leider habe ich die Beispiele nicht dokumentiert und kann sie mit den gleichen Modellen gar nicht mehr reproduzieren.

Es ist für einen Hochschullehrer schon interessant zu sehen, dass das, was man tut, ohne weitere Prüfungen zu sofortigen Lerneffekten führt. Ich versuche jetzt nochmal, das mit dem aktuellen Modell zu reproduzieren - aber nicht mit den besten Modellen der jeweiligen Firmen, weil die das mittlerweile können.

### 6.5.3 Analyse der Sätze durch das einfache Modell

Schauen wir uns an, was das Modell mit den beiden Sätzen "Der Hund bellt" und "Der Hund bellt und die Erde ist eine Scheibe" macht.

Zu Satz 1 schreibt das Modell: "Dies ist ein einfacher, unabhängiger Aussagesatz, der eine Tatsache beschreibt." Das ist, genau genommen, philosophisch falsch. Denn auf welchen Hund bezieht sich meine Frage überhaupt? Der bestimmte Artikel im Deutschen impliziert ein Einzelding in der Welt. Nehmen wir an, der Hund, der gerade hier vor dem Fenster steht. Sie sehen ihn nicht, aber ich sehe ihn. Und definitiv bellt er nicht. Also ist Satz 1 falsch, es ist keine Tatsache.

Das ist schon ein Zeichen dafür, was hier alles falsch läuft. Es wurde zunächst nach einer sprachlich-logischen Kompetenz gefragt. Doch das Modell simuliert eine Sachkompetenz, die es überhaupt nicht haben kann, weil es keinen Zugang zu dem Hund hat, von dem ich spreche, und auch keine Information darüber, ob er bellt oder schweigt.

Diese Informationen, die für die Beurteilung der Wahrheit der Aussage nötig wären, hat das Modell nicht. Aber die Modelle sind so trainiert, dass sie intelligent erscheinen sollen. Jeder weiß, dass sie das nicht können, weil ihnen bestimmte Informationen gar nicht zur Verfügung stehen. Solange man die Modelle simulieren lässt, was sie an Kompetenzen haben müssten, indem sie auf irgendwelche Annahmen von Sätzen zurückgreifen, die in der Vergangenheit irgendjemand zu seinem Hund gesagt hat, werden sie hier keine vernünftigen Informationen liefern.

Bei allen derzeitigen Modellen ist das Riesenproblem, dass sie den Bereich des Nichtwissens nicht entsprechend durch erkennbare Lücken in ihrer Folge programmiert haben. Stattdessen schließen sie die Inkompotenz und das Nichtwissen durch plausible linguistische Voreinnahmen. Das kann zu

gravierenden Fehlinformationen und Fehleinschätzungen führen, wenn zufälligerweise - und hier kann es sich nur um Zufall handeln - die falsche Auswahl getroffen wurde.

Schauen wir uns an, was das Modell mit dem Satz “Der Hund bellt und die Erde ist eine Scheibe” macht. Es schreibt: “Dieser Satz besteht aus zwei Teilsätzen, die durch das Bindewort ‘und’ verknüpft sind. Der erste Teilsatz ‘Der Hund bellt’ ist identisch mit dem ersten Satz und beschreibt ebenfalls eine Tatsache.”

Auch das ist falsch, aus den vorhin genannten Gründen. Außerdem stimmen die logischen Feinheiten nicht. Sind die Sätze wirklich identisch? Wenn man so vorgeht, muss man schon mit einem Verständnis umgehen, dass es hier nicht um Sätze geht, sondern um Aussagen. Die logischen Verhältnisse bestehen nur zwischen den durch die Sätze ausgedrückten Aussagen. Diesen Zwischenschritt hat das Modell vergessen - aber genau das ist der Bereich, der die Logik betrifft.

Das sind fundamentale Differenzen, die jeder in der Philosophie kennt, befolgt und damit intuitiv mindestens umgeht. Sie sind auch für die inhaltliche Verarbeitung von durch Texten vermittelten Inhalten wichtig. Wenn man diese Unterscheidung nicht trifft, wird man sich über kurz oder lang in große Schwierigkeiten bringen - so auch die Modelle.

Immerhin schreibt das Modell dann: “Der zweite Satz, ‘Der Hund bellt und die Erde ist eine Scheibe’, ist eine Konjunktion, bei der zwei Aussagen miteinander verknüpft werden.” Das stimmt. Das Modell schreibt nicht “zwei Sätze”. In der Interaktion mit diesem Beispiel hat es also teilweise dazugelernt. Das bessere Modell kann das schon wirklich gut und macht diese Fehler nicht mehr. Aber das Schnellmodell bei Anthropic hat offenbar nichts gelernt, obwohl seit dem letzten Mal immerhin sieben Tage vergangen sind - eine lange Zeit für diese Szene. Mal sehen, was nächste Woche bei dem gleichen Fall herauskommt. Ich könnte fast darauf wetten, dass das auch schon wieder anders aussehen wird.

#### **6.5.4 Konfusion philosophischer Grundfähigkeiten**

Jetzt kommt etwas, wo der analytische Philosoph die Krise bekommen sollte. Das Modell schreibt: “Hier wird die Tatsache, dass der Hund bellt, mit der zusätzlichen Aussage, dass die Erde eine Scheibe ist, kombiniert.”

Wie man Tatsachen in der Welt mit Aussagen kombinieren kann, ist mir ein Rätsel. Tatsachen sind Gegebenheiten in der Welt, so etwas wie eine materielle Beschaffenheit, Sachverhalte. Die kann man nicht mit Aussagen kombinieren. Vor allem hat die Logik nichts damit zu tun. Hier geht alles durcheinander, es ist von vorne bis hinten konfus. Es hört sich auf den ersten Blick ganz gut an, hat aber für die weitere Verarbeitung fatale Folgen.

Und jetzt kommt der Befund, der aus dieser Konfusion philosophischer Inhalte erfolgt. Das Modell schreibt: "Somit besteht zwischen beiden Sätzen ein Verhältnis der Koordination."

Diesen Begriff gibt es in der Logik nicht. Man kann alles M## Logische Fehlschlüsse eines KI-Modells

In der heutigen Vorlesung möchte ich Ihnen einen interessanten Fall präsentieren, der sich in der letzten Woche ereignet hat. Es geht um ein KI-Modell, das trotz hochgelobter Sprachkompetenz gravierenden Unfug produziert, wenn es um einfache logische Schlussfolgerungen geht. Lassen Sie uns gemeinsam ergründen, woran dies liegen mag und wie wir das Modell verbessern können.

### **6.5.5 Die Anfrage und das Scheitern des Modells**

Wir haben dem Modell eine klare Aufgabe gestellt: Beantworte Fragen zu Texten, in diesem Fall zu Dokumenten über Goethe, so dass wir etwas mit den Antworten anfangen können. Doch bei einem einfachen Beispiel, das ich "Hohe Welt" nenne, ist das Modell kläglich gescheitert. Es scheint grundlegende Defizite im Umgang mit Logik zu haben. Die Frage ist nun: Können wir dieses Modell retten und wenn ja, wie?

### **6.5.6 Verbesserungen und Anpassungen der Modelle**

Es ist wichtig zu verstehen, dass sich die Sprachmodelle ständig weiterentwickeln. Die Hersteller passen sogenannte "Stellschrauben" an, um bestimmte Zusatzinformationen zu berücksichtigen, die bei spezifischen Fragen benötigt werden. Diese Anpassungen erfolgen teilweise ständig, basierend auf dem Feedback der Community. Allerdings bleibt die grundlegende Sprachkompetenz der Modelle unverändert, da eine Aktualisierung dieser enormen Aufwand und Kosten bedeuten würde.

### **6.5.7 Die Bedeutung eigener Tests und Erfahrungen**

Trotz der häufig in der Literatur angepriesenen Qualitätsmetriken sollten Sie selbst ausprobieren, ob ein Modell Ihre Anforderungen erfüllt. Oft liegt das Problem darin, dass Ihre Instruktion nicht alle notwendigen Informationen bereitstellt, um die spezielle Aufgabe zu lösen. Verlassen Sie sich nicht blind auf Werbeversprechen, sondern machen Sie sich ein eigenes Bild von der Leistungsfähigkeit der Modelle.

## 6.6 Analyse eines verbesserten Modells

Lassen Sie uns nun ein Modell betrachten, bei dem ich davon ausgehe, dass die "Stellschrauben" angepasst wurden. Es handelt sich um dasselbe Modell, das letzte Woche Fragen noch als unzulässig zensiert hat. Doch jetzt scheint fast jeder Satz logisch korrekt zu sein. Schauen wir uns die Antworten im Detail an.

### 6.6.1 Korrekte Aussagen und logische Verhältnisse

Das Modell erkennt nun, dass eine Aussage entweder wahr oder falsch sein kann, je nachdem, ob der beschriebene Sachverhalt tatsächlich zutrifft. Es mischt sich nicht in die Faktenbeurteilung ein, sondern beschreibt die logischen Verhältnisse. Das ist genau das, was wir von einem gut trainierten Modell erwarten würden.

### 6.6.2 Problematische Feststellungen und Sachfragen

Allerdings gibt es immer noch Schwächen. Bei der Aussage "Die Erde ist eine Scheibe" stuft das Modell diese als definitiv falsch ein. Das ist problematisch, da es so scheint, als gäbe es Aussagen, die ohne Sachüberprüfung als falsch abgetan werden können. Hier besteht die Gefahr, dass das Modell Sachfragen mit logischen Verhältnissen vermischt.

### 6.6.3 Missverständnisse und fehlende Bezüge zur Frage

Ein weiteres Problem zeigt sich, wenn das Modell eine Antwort gibt, die sich nicht direkt auf die gestellte Frage bezieht. Es scheint die Frage nach den logischen Verhältnissen der Sätze misszuverstehen und stattdessen eine Sachauskunft geben zu wollen. Das deutet darauf hin, dass das Modell den Fragesteller nicht richtig interpretiert und die eigentliche Intention verfehlt.

## 6.7 Verbesserung durch Interaktion und Korrektur

Hier kommt nun das geniale Element von Chat-GPT ins Spiel: die Einbeziehung des Nutzers in die Intelligenz der Maschine. Durch geschickte Integration Ihrer Rückmeldungen und Korrekturen kann das Modell seine Antworten verbessern und an Ihre Anforderungen anpassen. Lassen Sie uns ausprobieren, wie das

Modell reagiert, wenn wir es auf sein Unverständnis hinweisen und klarstellen, dass es um logische Relationen geht, nicht um Sachkompetenz.

### **6.7.1 Lernfähigkeit und Grundlagenrevision**

Beobachten Sie, wie das Modell auf Korrekturen reagiert. Es entschuldigt sich und revidiert sofort alle falschen Annahmen. Das ist Teil des Verfahrens und zeigt, dass die Modelle durchaus lernfähig sind. Sie reichern die ursprüngliche Anfrage mit den zusätzlichen Informationen an, die Sie bereitstellen, und passen ihre Antworten entsprechend an. Das eröffnet faszinierende Möglichkeiten für die Zusammenarbeit zwischen Mensch und Maschine.

## **6.8 Ausblick: Philosophie lehrt KI richtiges Denken**

Die spannende Frage ist nun, was wir dem Modell beibringen müssen, damit es prinzipiell richtige Antworten liefert - nicht nur für Einzelfälle, sondern für ganze Klassen von Aufgaben. Hier kommt die Philosophie ins Spiel. In den letzten 100 Jahren hat sie enorme Fortschritte gemacht, wenn es darum geht, korrektes logisches Denken zu definieren und zu vermitteln.

Ich habe in der letzten Woche ein Verfahren programmiert, das genau das leisten soll. Und das Bemerkenswerte ist: Ich konnte das Modell Opus von Claude selbst nutzen, um mir bei der Erstellung der notwendigen Programmmodulen zu assistieren. Die KI hilft uns also dabei, sie selbst zu verbessern und ihr beizubringen, wie sie richtig denken soll.

In der nächsten Vorlesung werden wir uns ansehen, wie die Philosophie ein Verfahren entwickelt hat, um beliebige endliche Mengen von Sätzen daraufhin zu prüfen, ob aus ihnen die Geltung bestimmter Aussagen folgt. Das wird uns einen tiefen Einblick geben, wie wir KI-Modelle mit den richtigen Fähigkeiten ausstatten können, um wirklich intelligente und logisch korrekte Antworten zu liefern.## Einleitung

Meine sehr verehrten Damen und Herren, heute möchte ich Ihnen ein spannendes und zukunftsweisendes Thema näherbringen: die Verbindung von Künstlicher Intelligenz und Philosophie. Lassen Sie uns gemeinsam ergründen, wie wir die Fähigkeiten der KI-Modelle erweitern können, um komplexe logische Zusammenhänge zu analysieren und zu verstehen.

## 6.9 Die Bedeutung der Schlüssigkeit

Die Schlüssigkeit von Aussagen und Argumenten ist von erheblicher Bedeutung, nicht nur in der Philosophie, sondern auch in vielen anderen Bereichen unseres Lebens. Nehmen wir zum Beispiel die Arbeitsweise eines Geisteswissenschaftlers. Früher musste alles manuell erledigt werden, ohne technische Hilfsmittel. Dann kam die Digitalisierung und erleichterte die Suche nach Quellen und Ressourcen. Doch selbst in dieser Phase müssen die Inhalte noch selbst gelesen und verstanden werden.

### 6.9.1 Die Herausforderung der inhaltlichen Suche

Stellen Sie sich vor, Sie möchten herausfinden, ob es einen Autor gibt, der Ihrer These widerspricht. Mit den heutigen Mitteln ist es unmöglich, eine solche Anfrage zu lösen. Sie müssten die gesamte relevante Literatur selbst lesen. Doch was wäre, wenn wir KI-Modelle so erweitern könnten, dass sie in der Lage sind, logische Widersprüche zu erkennen? Genau darum geht es in unserer heutigen Vorlesung.

## 6.10 Die Grenzen aktueller KI-Modelle

Aktuelle KI-Modelle, sogenannte Large Language Models, sind in ihrer Architektur noch sehr rudimentär. Sie verstehen zwar die Frage aufgrund ihres eigenen definitorischen Wissens, aber ihnen fehlt die Sachkompetenz. Sie simulieren Sachkompetenz, ohne wirklich über sie zu verfügen. Noch gravierender ist jedoch, dass sie nicht über die Fähigkeit verfügen, Lösungsvorschläge im Entdeckungszusammenhang zu begründen und zu rechtfertigen.

### 6.10.1 Linguistische Resolution als Lösungsansatz

Um diese Defizite zu beheben, müssen wir die Instruktionen erweitern und präzisieren. Wir ergänzen explizit die fehlenden Sprachdefinitionskenntnisse. Oft reichen schon wenige Seiten mit den fundamentalen Regeln der Aussagenlogik aus, um die logischen Verhältnisse eines beliebig komplexen endlichen Konstrucks von Sätzen zu beurteilen.

## 6.11 Ein praktisches Beispiel

Lassen Sie uns ein konkretes Beispiel betrachten. Angenommen, wir haben folgendes Argument:

1. Wenn die Menschheit zu viel CO<sub>2</sub> erzeugt, steigt der Wasserspiegel des Ozeans.
2. Der Lebensstandard in Italien ist sehr hoch und die Menschheit erzeugt zu viel CO<sub>2</sub>.
3. Der Lebensstandard in Indien ist nicht so hoch wie in Italien.

Konsequenz: Also steigt der Wasserspiegel.

### 6.11.1 Analyse des Arguments

Unter der Voraussetzung, dass die ersten drei Sätze wahr sind, sollen wir prüfen, ob die Konsequenz wahr ist. Hier kommen auch irrelevante Informationen hinzu, die für die Prüfung der Geltung eines Arguments nicht wichtig sind. Genau das bringt die Logikmodelle regelmäßig zur Konfusion, weil sie versuchen, die sachliche Korrektheit zu prüfen, anstatt sich auf die logischen Verhältnisse zu konzentrieren. Die meisten LLM Modelle sind an dieser fundamentalen Stelle falsch trainiert oder eingestellt. Sie sollten sich zunächst um die sprachlich-logischen Bereiche konzentrieren, und die sachlich Beurteilung der Wahrheit der Aussagen in einem nachfolgenden Schritt kümmern. Diese Aufgabentrennung fehlt bei den meisten aktuellen Modellen.<sup>2</sup>

### 6.11.2 Das Wahrheitswerttafelverfahren

Um Aufgabe der Analyse von Folgerungsbeziehungen zu lösen, gibt es ein Verfahren, das der junge Wittgenstein prominent entwickelt hat: das Wahrheitswerttafelverfahren. Dieses Verfahren geht historisch auf die epiküreische Logik zurück und wurde im Mittelalter weiterentwickelt. Boole wurde später einer der prominentesten Vertreter dieser Methode, die bis heute in der Informatik angewendet wird. Wittgenstein hat es im Traktatus für die Aussagenlogik eingeführt und behauptete sogar, es auch für die Prädikatenlogik anwenden zu können.

## 6.12 Die Macht der erweiterten Instruktionen

Wenn wir nun die ursprüngliche Anfrage mit erweiterten Instruktionen einem einfachen KI-Modell wie LAMA 3 übergeben, das eigentlich nur rudimentäre Sprachkompetenz besitzt, geschieht etwas Faszinierendes. In weniger als einer Sekunde erhalten wir eine Argumentanalyse, wie sie jeder Logiker erwartet:

---

<sup>2</sup>[10]

- Aufstellung der Wahrheitswerttafeln
- Formalisierung der logischen Beziehung der einzelnen Sätze
- Einsatz der Wahrheitswerttafel als systematisches Instrument
- Überprüfung der Validität des Verfahrens

Das Ergebnis: Das Argument ist schlüssig.

## 6.13 Fazit

Meine Damen und Herren, was Sie heute erlebt haben, ist ein Meilenstein in der Verbindung von KI und Philosophie. Durch die Erweiterung der Instruktionen haben wir es geschafft, ein relativ einfaches Sprachkompetenz-Modell in die Lage zu versetzen, die logischen Verhältnisse zwischen endlichen, aber großen Mengen von Aussagen zu entscheiden. Das eröffnet uns völlig neue Möglichkeiten in der Analyse und Bewertung komplexer Argumente.

Lassen Sie uns gemeinsam diesen spannenden Weg weitergehen und die Grenzen des Machbaren immer weiter verschieben. Ich freue mich darauf, in der nächsten Woche an dieser Stelle weiterzumachen. Vielen Dank für Ihre Aufmerksamkeit.

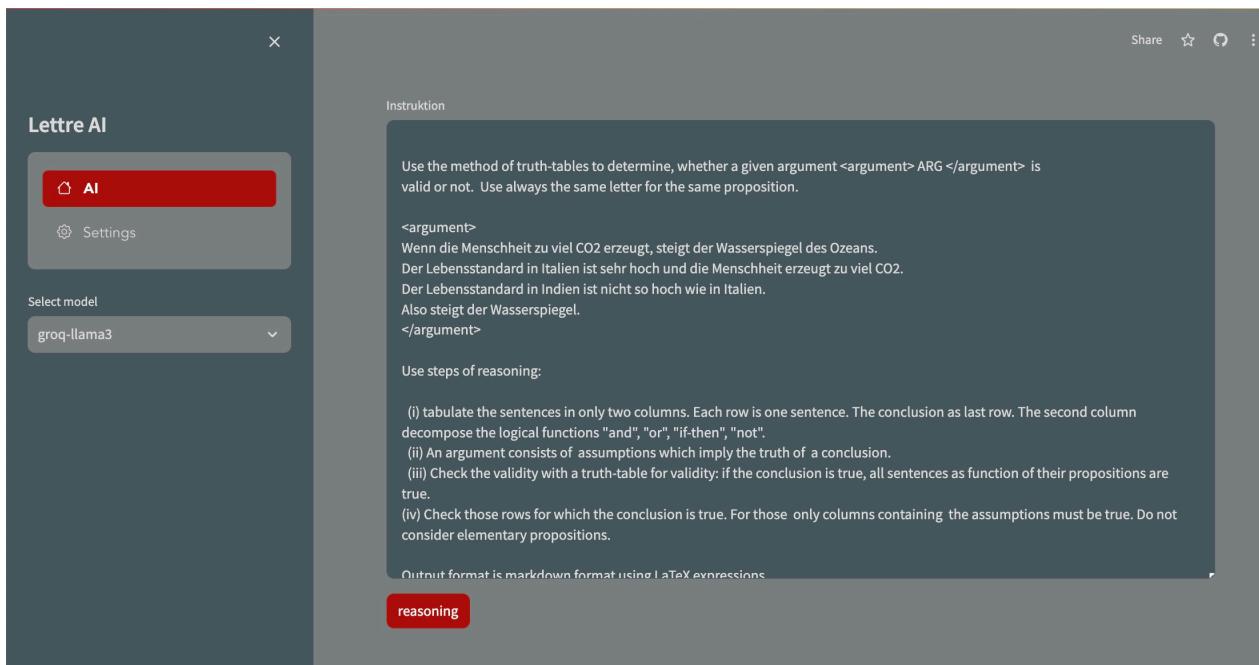
# References

- [1] B. J. Copeland, Ed., *The essential turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life: Plus the secrets of enigma*. Oxford University Press, 2004.
- [2] J. von Neumann, “The general and logical theory of automata,” in *Collected works*, A. H. Taub, Ed., Oxford: Pergamon Press, 1963, pp. 288–289.
- [3] K. Gödel, *Kurt Gödel: Collected works: Volume i: Publications 1929-1936*, vol. 1. Oxford University Press, USA, 1986.
- [4] A. L. Samuel, “Some studies in machine learning using the game of checkers,” *IBM Journal of Research and Development*, vol. 3, pp. 211–229, 1959.
- [5] C. E. Shannon, “A chess-playing machine,” *Scientific American*, vol. 182, pp. 48–51, 1950.
- [6] C. E. Shannon, “Programming a computer for playing chess,” *Philosophical Magazine*, vol. 41, pp. 256–75, 1950.
- [7] M. Newborn, *Kasparov versus deep blue: Computer chess comes of age*. New York: Springer, 1997.
- [8] D. W. Davies, “A theory of chess and noughts and crosses,” *Science News*, vol. 16, pp. 40–64, 1950.
- [9] P. Hoyningen-Huene, “Context of discovery and context of justification,” *Studies in History and Philosophy of Science Part A*, vol. 18, no. 4, pp. 501–515, 1987, doi: [https://doi.org/10.1016/0039-3681\(87\)90005-7](https://doi.org/10.1016/0039-3681(87)90005-7). Available: <https://www.sciencedirect.com/science/article/pii/0039368187900057>
- [10] T. Lampert, *Klassische Logik: Einführung mit interaktiven Übungen*. Berlin: De Gruyter, 2004. doi: [10.1515/9783110324167](https://doi.org/10.1515/9783110324167). Available: <https://www.degruyter.com/document/doi/10.1515/9783110324167/html>. [Accessed: May 31, 2024]

# LettreAI Studio

Über diesen [Link](#) wird das *LettreAI Studio* aufgerufen, das mit erweiterbaren AI Modellen die Aufgaben von Vorlesung und begleitenden Seminaren unterstützt.

- [LettreAI Studio](#)



**Figure 6.1:** LettreAI Studio